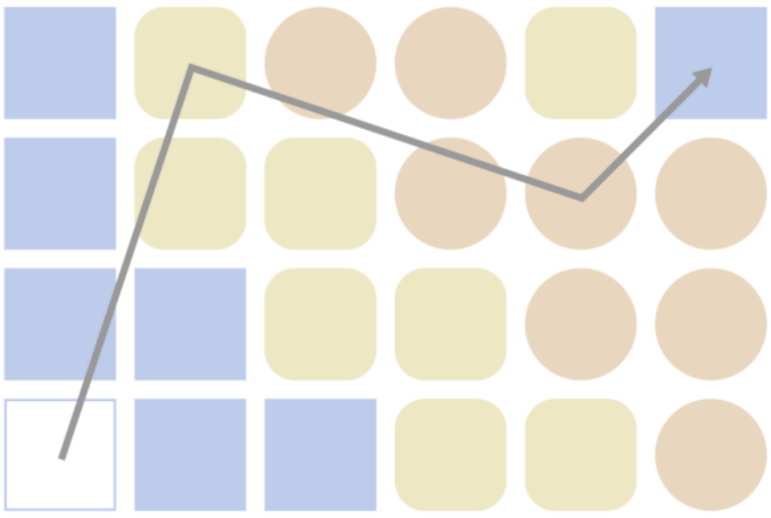
Congressi
Stefano Franscini
Swiss Federal Institute
of Technology Zurich

Biological systems: from first principles to data-driven modelling and back

Ascona, March 27 - April 1 2022



Organised by Niko Beerenwinkel, Peter Bühlmann, Wolfgang Huber

The data revolution in biology of the last two decades, driven by progress in genomics, imaging and other high-throughput technologies, and by progress in computing, has led to breathtaking advances in our ability to understand biological systems through inductive means, viz. pattern recognition, machine learning, and statistical modelling. On the other hand, theoretical biology is an established field following a deductive approach, where inspired by the success of physics, the dynamics of biological systems are described by mathematical models. Arguably, theory has been outpaced by the rapid developments in big-data biology. This workshop aims to bring together scientists from both sides, and in particular those working on the interface between them, to fuel a renewed effort in closing the loop between induction and deduction which is so essential for scientific progress. Specifically, we will (i) explore recent advances and open problems in mathematical modeling of biological systems comprising both approaches based on first principles and on large high-dimensional data sets; (ii) identify opportunities and challenges for these two approaches to benefit from each other or even to converge; and (iii) facilitate meaningful interactions between engineering, biomedical, and quantitative researchers of both camps.

Sponsors


Congressi
Stefano Franscini
Politecnico federale
di Zurigo


SIB
Swiss Institute of
Bioinformatics


Roche


NOVARTIS


Strategic Focus Area
Personalized Health
and Related Technologies

Venue

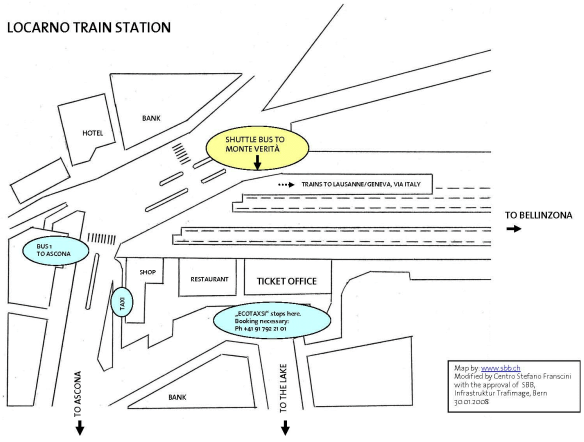
Monte Verità
Via Collina 84
CH-6612 Ascona
tel. +41 91 785 40 40

About the Congressi Stefano Franscini (CSF)

The Congressi Stefano Franscini (CSF) is the international conference centre of the Swiss Federal Institute of Technology (ETH) in Zürich, situated in the south of Switzerland (Canton Ticino) at Monte Verità. It has been named after the Federal Councillor Stefano Franscini, a native of Ticino who, in 1854, played an important part in establishing the first Federal Institute of Technology in Switzerland, ETH Zürich. Every year, the centre hosts 20 - 25 conferences organised by professors working at Swiss universities and concerning all disciplines (sciences and humanities) taught at academic level. The centre is also open to the local population with a regular program of public events (lectures, concerts, films, etc.) organised in the context of its international conferences and/or Monte Verità's cultural programme.

Shuttle service from Locarno Station

A free 9-seater shuttle bus to Monte Verità leaves from Locarno railway station Sunday April 27 at the following times: **13.30**; **14.30**; **15.30**; **16.30**; **17.30**. The meeting point is on the right side of the train platforms in Locarno (see image).



Keynote lectures

Epigenomic liquid biopsy - from basic science to translational applications

Nir Friedman

Hebrew University of Jerusalem

Collective behavior in interacting cell systems: from first principles to mathematical analysis and back

Andreas Deutsch

TU Dresden

Collective dynamics in multicellular systems plays key roles in biological development, regeneration, and disease. It can be studied with on- and off-lattice agent-based models. Here, I introduce an on-lattice, agent-based model, which I call biological lattice-gas cellular automaton (BIO-LGCA). I introduce elementary BIO-LGCA models of fundamental cell interactions, which may be combined in a modular fashion to analyse complex multicellular phenomena. Furthermore, I provide examples for the mathematical mean-field analysis of specific BIO-LGCA models, which allow to explain various kinds of collective behaviour. The first example demonstrates how to identify the mechanisms underlying the formation of collective invasion patterns recently discovered in breast cancer populations. The second example sheds light on the emergence of genetic heterogeneity due to

cellular interactions in growing tumors.

How personalised is your immune repertoire

Aleksandra Walczak

ENS Paris

Immune repertoires provide a unique fingerprint reflecting the immune history of individuals, with potential applications in precision medicine. Can this information be used to identify a person uniquely? If it really is a personalised medical record, can it inform us about the outcomes of a COVID-19 infection? I will explore how learning about randomness can help us read this information we all carry within us.

Algorithm validation for data science

Joachim Buhmann

ETH Zürich

Data Science (DS) algorithms interpret outcomes of empirical experiments with random influences. Often, such algorithms are cascaded to long processing pipelines especially in biomedical applications. The validation of such pipelines poses an open question since data compression of the input should preserve as much information as possible to distinguish between possible outputs. Starting with a minimum description length argument for model selection we motivate a localization criterion as a lower bound that achieves information theoretical optimality. Uncertainty in the input causes a rate distortion tradeoff in the output when the DS algorithm is adapted by learning. We present design choices for algorithm selection and sketch a theory of validation. The concept is demonstrated in neuroscience applications of diffusion tensor imaging for tractography and brain parcellation.

Instrumental variables in sparse and synamical settings

Jonas Peters

University of Copenhagen

Exogenous heterogeneity, for example in the form of instrumental variables, can help us learn a system's underlying causal structure and predict the outcome of unseen intervention experiments. In this talk, we discuss this idea in a setting in which the causal effect from covariates on the response is sparse and in a setting, where the variables follow a time dependence structure. If time allows, we also briefly discuss what can be done when identifiability conditions are not satisfied.

Metabolic fitness landscapes predict the evolution of antibiotic resistance

Fernanda Pinheiro

Human Technopole

Bacteria evolve resistance to antibiotics by a multitude of mechanisms. A central, yet unsolved question is how resistance evolution affects cell growth at different drug levels. Here, we develop a fitness model that predicts growth rates of common resistance mutants from their effects on cell metabolism. The model maps metabolic effects of resistance mutations in drug-free environments and under drug challenge; the resulting fitness trade-off defines a Pareto surface of resistance evolution. We predict evolutionary trajectories of growth rates and resistance levels, which characterize Pareto resistance mutations emerging at different drug dosages. We also predict the prevalent resistance mechanism depending on drug and nutrient levels: low-dosage drug defence is mounted by regulation, evolution of distinct metabolic sectors sets in at successive threshold dosages. Evolutionary resistance mechanisms include membrane permeability changes and drug target mutations. These predictions are confirmed by empirical growth inhibition curves and genomic data of Escherichia coli populations. Our results show that resistance evolu-

tion, by coupling major metabolic pathways, is strongly intertwined with systems biology and ecology of microbial populations.

Reconstructing, tracking, and predicting viral spread and evolution

Richard Neher

University of Basel & SIB Swiss Institute of Bioinformatics

The SARS-CoV-2 pandemic has resulted in unprecedented genomic and epidemiological surveillance as well as virological characterization in record time. The volume of data and requirement for rapid feedback to decision makers required continuous data analysis, method development, and infrastructure up-scaling. I will discuss what we can learn from such data about viral evolution, how these insights can inform response, and how similar approaches could inform the management of other infectious disease.

Contributed talks

Modelling of epigenetic clock based on nonlinear methylation trajectories

Alena van Bömmel

Leibniz Institute on Aging – Fritz Lipmann Institute (FLI)

Age-related DNA methylation changes in CpG dinucleotides have been shown to accurately estimate the underlying chronological age, a phenomenon known as epigenetic clock. Although the human genome contains more than 20 million CpG dinucleotides, only a small fraction is necessary for forming the epigenetic clock. To date, modelling of epigenetic clocks focused on the linear relationship between the DNA methylation status and age. DNA hydroxymethylation (5-hmC), is a recently identified type of DNA modification in which the hydrogen atom at the C5-position in cytosine is replaced by a hydroxymethyl group, and whose role as an epigenetic marker have not been yet fully understood. Here, we present modelling of epigenetic clock based on DNA hydroxymethylation measurements in young, middle age and geriatric mice that focus on nonlinear trajectories in aging processes. First, we identify CG clusters with coordinated aging behavior across samples. Then, we apply sparse group lasso models to construct a novel epigenetic clock based on methylation patterns in clusters with nonlinear trajectories. Our novel epigenetic clock has a high correlation (r=0.95) with the chronological age and shows the potential of the DNA hydroxymethylation to predict age. Furthermore, we identify CG clusters with distinct behavior in young, middle age and geriatric mice that are predictive for aging. In conclusion, we identified epigenetic biomarkers of aging based on a novel method that could provide more insights of the underlying molecular processes of aging.

EPISPOT: an epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies

Hélène Ruffieux

MRC Biostatistics Unit, University of Cambridge

We present EPISPOT, a scalable hierarchical framework which exploits large panels

of epigenetic annotations as variant-level information to enhance molecular quantitative trait locus (QTL) mapping. While existing epigenome-driven approaches can only model cis regulation or genetic association with clinical phenotypes, EPISPOT accommodates functional information for both cis and trans action, including QTL hotspot effects. Specifically, we consider a series of parallel regressions combined in a hierarchical manner to flexibly accommodate high-dimensional responses (molecular outcomes) and predictors (genetic variants), thereby allowing information-sharing across outcomes and variants. We also directly model the propensity of variants to be hotspots, i.e., to remotely control the levels of many gene products, via a dedicated top-level representation. We implement a variational expectation-maximisation algorithm augmented with simulated annealing schemes to enhance exploration of highly multimodal spaces and allow simultaneous analysis of thousands of samples, responses, predictors and predictor-level annotations. Hence, our approach effectively couples joint QTL analysis and hypothesis-free selection of annotations which directly contribute to the QTL effects; this unified learning boosts statistical power and helps shed light on the regulatory basis of the uncovered QTL hits. EPISPOT therefore marks a step forward to improving the challenging detection and functional interpretation of trans effects. We illustrate the advantages of EPISPOT in simulations emulating real-data conditions and in a monocyte expression QTL study, which confirms known hotspots and reveals new ones, as well as plausible mechanisms of action.

Multi-omics and artificial intelligence to fight cancer

Emmanuel Barillot

Institut Curie

We now have many modalities to explore the tumor and its environement: many types of omics, imaging, clinical records... One burning question is to convert this heaps of data into knowledge useful in clinics. This requires ad hoc algorithms which are capturing the underlying biological nature of the data while being computationally efficient which means keeping a reasonable level of complexity. I will present some examples of our research in that direction: how to reconstruct evolutionary path from clinical data? how to integrate multi-omics to gain biological insights and clinical perspectives in a devastating pediatric disease: medulloblastoma how to model the spatial dimension of the tumor, a critical though neglected aspect of cancer.

Mapping single-cell transcriptomes to copy number evolutionary trees

Pedro Ferreira

CBG, BSSE, ETH Zürich

Cancer arises and evolves by the accumulation of somatic mutations that provide a selective advantage. The interplay of mutations and their functional consequences shape the evolutionary dynamics of tumors and contribute to different clinical outcomes. In the absence of scalable methods to jointly assay genomic and transcriptomic profiles of the same individual cell, the two data modalities are usually measured separately and need to be integrated computationally. Here, we introduce SCATrEx, a statistical model to map single-cell gene expression data onto the evolutionary history of copy number alterations of the tumor. SCATrEx jointly assigns cancer cells assayed with scRNA-seq to copy number profiles arranged in a copy number aberration tree and augments the tree with clone-specific clusters. Our simulations show that SCATrEx improves over both state-of-the-art unsupervised clustering methods and cell-to-clone assignment methods. In an application to real data, we observe that SCATrEx finds inter-clone and intra-clone gene expression heterogeneity not detectable using other integration methods. SCATrEx will allow for a better understanding of tumor evolution by jointly analysing the genomic and transcriptomic changes that drive it.

The spatial distribution of clonal dynaics in the earliest stages of skin carcinogenesis

Pablo Baeza Centurion

German Cancer Research Center (DKFZ)

In pre-malignant tissue before tumours are even formed, cells accumulate mutations that generate clones if they are selected for. However, the processes that underlie mutational selection in such an environment remain poorly understood. To address this, we used one of the most common mouse models employed to study the development of skin can-

cer: two-stage chemically-induced carcinogenesis with DMBA and TPA. By extracting hundreds of biopsies from a 2 squared centimetre area of mouse skin, we generated spatially aware whole-exome and bulk RNA sequencing data, in addition to single-cell RNA sequencing data.

The proportion and spatial distribution of different cell types changed drastically upon treatment, with certain cell types being associated with the appearance of premalignant papillomas and allowing us to distinguish between populations of anti-tumour and tumour-promoting T cells. We analysed the spatial distribution of selection before and after treatment. This revealed that selection is very unevenly distributed in an area of mouse skin, with regions under strong selection found next to regions whose clones are growing neutrally. Finally, we analysed the distributions of clone sizes in each biopsy to identify areas of the skin containing "pre-papillomas" – clones that are growing more than expected but which derive from biopsies that macroscopically do not appear different from normal skin. Therefore, our data reveal genomic features that may allow the early detection of pre-malignant clones.

Cell cycle phase inheritance models to reveal biological oscillators that drairve the cell cycle

Fern A. Hughes, Alexis R. Barr, Philipp Thomas

Imperial College London

Advances in time-lapse microscopy mean individual cells can be tracked as they move through the cell cycle and their lineage information obtained. In the literature, many interesting results have been revealed by analysing the correlation in interdivision time of various family pairs of cells.

We introduce a general stochastic matrix model of abstract cell-cycle factors that are inherited from the mother to daughter cell. The model can be parameterised from measured correlations alone, and contains common models of cell cycle, replication and size control as special cases.

Using Bayesian inference, we fit the model to six lineage tree datasets of bacteria and mammalian cells from available literature. Our model gives a consistently good fit to the data, despite the model parameters being unidentifiable. This suggests that using interdivision time data alone is not sufficient to identify precise cell cycle control mecha-

nisms. However, analysing the posterior distributions reveals oscillations driving the cell cycle. The frequencies of these oscillations can be attributed to underlying biological oscillators, such as circadian rhythm. This insight helps us to understand the factors that affect cell-to-cell variability in interdivision time.

Agent-based branching process

Francesco Puccioni

Imperial College London

Quantifying cellular growth is crucial to understanding the dynamics of cell populations such as microbes and cancer cells. Standard growth curves of batch cultures are characterized by a lag-phase, exponential growth, and a stationary phase; however, at the single-cell level, growth varies drastically from cell to cell and from lineage to lineage. The sources of this heterogeneity can be found in the stochasticity of cell cycle duration, heterogeneous adaption to changing environments, and demographic noise due to cell death that affects population structure.
I propose a novel stochastic age-dependent model where the cells are represented by agents who divide and die in response to an internal stochastic state which is continuous and increases with time. While such agent-based models are usually only amenable to simulations, we show that the population structure can be characterized by a functional master equation cast as a functional derivative equation. We manipulated the master equation to recall a novel integral renewal equation for the generating functional. Compared to the classic results about renewal theory, the latter equation takes a step further. De facto, it provides a wide stochastic description for continuous structured populations driven by birth and death reactions.
In the first part of the project, we recalled the analytic expressions for all the stochastic moments of the abundances, as for the internal traits density and lineages distribution. We also obtained analytical and numerical solutions for the extinction probability and first passage times distribution which allow describing exactly the extinction probability (and its time distribution) for any internal state range of values.
In conclusion, we also developed a novel stochastic simulation method based on the rejection-acceptance theory to simulate the evolution of continuous age-structured populations in order to substantiate the analytical and numerical results.

Genotype-phenotype relations of realistic 3-node networks

Eve Tasiudi

Department of Biosystems Science and Engineering, ETH Zürich

A fundamental problem in biology is the mapping between the genotype and the phenotype. A single phenotype can be the result of multiple genotypes that differ by small mutational changes. This set of multiple genotypes, called a genotype network, promotes evolutionary innovation by enabling the exploration of distinct and novel phenotypes. Concerning the genotype networks of gene regulatory networks (GRNs) there have been three types of modelling methods: i) abstract regulatory networks, ii) Boolean networks and iii) network theory analysis. These approaches help to interpret properties such as redundancy, bias, robustness and evolvability on large, interconnected genotype networks. However, the models are too abstract to infer the mechanisms that generate a specific phenotype. In addition, direct experimental evidence of the existence of genotype networks in GRNs and their role in evolutionary innovation has remained elusive. Combining ordinary differential equation models with data from twenty synthetic constructs of GRNs in Escherichia coli, we create the full set of genotype networks. Through an iterative process of estimation and validation, this framework allows us to identify interfacing genotype networks and determine the mechanisms that alter the observed phenotype. As a starting GRN, we chose a type 2 incoherent feed-forward loop (IFFL-2); its observed phenotype is a stripe of gene expression (high-low-high) in response to a chemical gradient. The small mutational changes that we implement are either qualitative (modifying the model topology) or quantitative (varying the inhibition strength). An exhaustive analysis of the map and its related mechanisms show that altering the inhibition strength in the same model topology is sufficient to alter the observed phenotype. Our work, apart from providing evidence on the existence of genotype networks for GRNs, pinpoints the components that take part in realizing the observed phenotype.

Analysis of multi-condition single-cell datasets using latent interaction models

Constantin Ahlmann-Eltze

EMBL Heidelberg

Complex modern biological datasets encompass variability from known and unknown covariates—for example, cells in multi-condition single-cell datasets group by treatment condition and by cell type. One motivation to conduct single-cell sequencing, in contrast to bulk, is the ability to find cell type-specific treatment effects. However, current tools either require the annotation of cell types in all conditions, which is only feasible for data with a limited number of conditions and clear cell type boundaries, or apply non-linear data integration methods which complicate downstream interpretation and are known to over-adjust and obscure biologically relevant differences.
Here, we present a novel statistical framework to infer the latent structure of such datasets and quantify the cell type-specific treatment effects. Our approach extends the classical multivariate linear model framework by including interaction terms between selected known and inferred covariates. Our model describes the data with three terms: (1) the overall treatment effect, (2) the latent structure that is shared across conditions (e.g., common cell types), and (3) the cell type-specific treatment effects. Employing interaction terms means that our model is not just limited to discrete latent structures but can naturally handle gradual changes in the latent space like developmental trajectories. Furthermore, the linearity of our model means that we can interpret the coefficients of the model directly in terms of their effect on gene expression. We establish the general principles of our model on simulated data and demonstrate its utility by re-analyzing several publicly available single-cell datasets.

Gaussian mixture modelling of V gene diversity distribution reveals different levels of heterogeneity in human naïve and memory CD4 T cells in a-bomb survivors

Justyna Mika

Silesian University of Technology

T lymphocytes play an essential role in the defense against pathogens and cancers through their clonally distributed T cell receptors (TR). TR genes are assembled from discrete V, D and J segments in developing lymphocytes. The V genes are part of the variable TR domain, and their distribution is often used as a measure of T cell diversity.
Frequency of 11 V genes was measured by flow cytometry in 674 human blood samples obtained from A-bomb survivors about 50 years after exposure (doi:10.1046/j.1365-2141.2003.04520.x). To describe the diversity of V genes in naïve and memory cells, Pielou's index was calculated, which ranges from 0 to 1 indicating heterogeneous and uniform distribution of V genes respectively. Kolmogorov-Smirnov (KS) test was applied to compare the distribution of V gene diversity between naïve and memory cells. To distinguish groups of donors with different levels of diversity, Gaussian Mixture Modeling (GMM) was used with Bayesian Information Criterion for model selection.
Memory cells diversity is lower than that of naive cells (KS p-val $< 1e-6$). The distribution of V gene frequency is more heterogeneous in memory cells, meaning that some V genes are over- or under- represented. Applying GMM models on naïve and memory cells diversity distributions, resulted in finding groups of low and high diversity. The mean diversity of the high diversity component is the same in naïve and memory cells, but the contribution of this component is much more important in naïve than in memory cells, with a respective weight 0.8282 vs 0.5809. Thus, a significant subset of donors has a V gene repertoire more heterogenous in memory cells than in naïve cells.
The work was supported by European Social Fund grant POWR.03.02.00-00-I029 and Silesian University of Technology grant for Support and Development of Research Potential.

Towards a truly integrative systems immunology

Fridolin Gross

ImmunoConcept, University of Bordeaux

Interestingly, big data and models have shown very different levels of success in biological practice: while experimentalists routinely rely on the tools and results provided by bioinformatics, they appear in general much more reluctant to integrate methods of computational modeling into their research. This phenomenon is particularly pronounced in the field of immunology. There seems to be a need, voiced in many review articles, for a broader vision of "systems immunology" that uses the conceptual resources of both bioinformatics and computational modeling. However, while a cursory glance at the scientific literature shows that the notion of "systems immunology" in analogy to "systems biology" is quite popular among immunologists, it appears that in actual research contexts this notion is almost exclusively understood in terms of the integration of high-throughput methods with more traditional experimental methods. The results of computational models, by contrast, are rarely taken up by mainstream journals. Thus, the question arises what impedes the progress of modeling in immunology. Is the problem due to an aversion to mathematics on the part of experimental immunologists? Or are there more substantive reasons that can explain the challenges in bringing together informal reasoning and computational methods in certain research contexts? Should an incorporation of modeling be encouraged (and if so with which goals?), or are immunologists better off following their established ways of thinking? I propose a conceptual framework that helps to understand these differences in the success of integration. This framework conceives of the impact of computational methods in biology as a change at the level of cognitive practices from "informal" to "formal" strategies. I offer a philosophical analysis of these concepts that is informed by cognitive science and by a detailed investigation of representative historical and contemporary case studies.

Inferring context-specific signalling states using large-scale CRISPR-KO screens

Paula Weidemüller

EMBL-EBI

Large-scale CRISPR knockout screens covering a wide array of cell lines, for example DepMap and Project Score, revealed that a lot of genes are context-essential. The challenge is to systematically identify and define those contexts and how they affect gene interactions and cell signalling. Using CEN-tools (https://doi.org/10.15252/msb.20209698), we tried to identify context-specific parts of the cell signalling network. Inferring the context-dependent signalling state is crucial to understanding cellular phenotype and behaviour. Knowing the specific signalling network wiring can aid in predicting the cellular response to changes in environment e.g., treatment with a drug. To this end, we built a linear model to disentangle the tissue-specific effects of different driver gene mutations on gene essentiality. Contrary to expectations, we observe that the genes essential in the context of a cancer driver mutation are highly tissue specific. By mapping those genes onto a protein-protein interaction network we were able to extract context-specific essential signalling modules. These findings could lend insights into why drugs targeting the same cancer driver are effective in cancer of one tissue type but not in another. Furthermore, they can help identifying synthetic lethalities to aid drug development for targeted therapy.

Improving tumour predictions: a bayesian combination of dynamic modelling and machine learning approach

Haralampo Hatzikirou

Khalifa University

In clinical practice, a plethora of medical examinations are conducted to assess the state of a patient's pathology producing a variety of clinical data. However, exploiting these data faces the following challenges: (C1) we lack the knowledge of the mechanisms involved in regulating these data variables, and (C2) data collection is sparse in time since it relies on patient's clinical presentation. (C1) implies that only a small subset of the

relevant variables can be modeled by virtue of mathematical modeling. This limitation allows models to be effective in analyzing the qualitative dynamics of the system, but limits their predictive accuracy. On the other hand, statistical learning methods are well-suited for quantitative reproduction of data, but they do not provide mechanistic understanding of the investigated problem. Moreover, due to (C2) any algorithm is challenged in learning the corresponding disease dynamics. Herein, we propose a method, based on the Bayesian coupling of mathematical modeling and machine learning (BaM3), aiming at improving individualized predictions by addressing the aforementioned challenges. As a proof of concept, we evaluate the proposed method on a synthetic dataset for brain tumor growth and analyze its performance in predicting two major clinical outputs, namely tumor burden and infiltration. The BaM3 method results in improved predictions in almost all simulated patients, especially for those with a late clinical presentation. In addition, we test the proposed methodology in two settings dealing with real patient cohorts. In both cases, namely cancer growth in chronic lymphocytic leukemia and ovarian cancer, BaM3 predictions show good agreement with reported clinical data.

Separable effects

Mats Stensrud

The École polytechnique fédérale de Lausanne (EPFL)

Suppose the causal effect of exposure A on outcome Y is established, e.g. in a biological experiment or a randomized clinical trial. Then, researchers are often interested in understanding how the effect of A on Y is mediated through other variables M. Identification of such mediation effects is a popular but controversial topic in the causal inference literature.

In this talk, I will present recent work on the separable effects, a new type of causal estimands in mediation settings, inspired by a seminal exposure decomposition idea of Robins and Richardson (2010). I will give criteria that allow different interpretations of the separable effects and present identification conditions that can be evaluated in causal graphs. Furthermore, I extend the definition of separable effects to settings where investigators are interested in exposure effects conditional on a post-exposure variable.

The separable effects are compelling alternatives to existing estimands, such as natural direct and indirect effects (e.g. advocated by Pearl) and principal stratum effects (e.g.

advocated by Rubin): in particular, the separable effects are defined in an identifiable subgroup of the population and can be identified under assumptions that are empirically verifiable in a future experiment (unlike existing estimands, cross-world assumptions are avoided). To illustrate the new estimands, I will present analyses from three biomedical experiments.

Exploring transcription variability along the cell cycle with metabolic labelling data and stochastic modelling

Dimitrios Volteras

Imperial College London

Single-cell transcriptomics technologies are central in studying stochasticity in heterogeneous cell populations. However, scRNA-seq data only reveal a static snapshot of single-cell gene expression profiles and do not directly capture dynamics of transcription. Recently developed metabolic labelling protocols in single-cell transcriptomics aim to monitor transcriptional dynamics by adding temporal resolution to scRNA-seq data. We are focusing on one of these protocols, named scEU-seq [1], that combines metabolic labelling of spliced and unspliced transcripts with cell-cycle ordering information. Analysis of scEU-seq data allows us to record mean changes and variability in gene expression levels along the cell cycle. We have developed a stochastic model of mRNA kinetics to be integrated with these data, that captures bursty synthesis, splicing and degradation of labelled and unlabelled transcripts, driven by cell-cycle dependent rate parameters. The model reveals that cell-cycle modulation of parameters such as burst size or burst frequency can explain the observed changes in expression of particular cell-cycle regulated genes. The objective of this study is to create a statistical inference framework to fit the model to the observed mean and noise trajectories on a genome-wide level. This framework can potentially elucidate how mechanistic parameters of gene expression are regulated by the cell cycle as well as provide links between stochasticity of bursty gene expression and cell-cycle variability.

Reference:

[1] Nico Battich, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S. Baron, Marvin E. Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. Sequencing

metabolically labeled transcripts in single cells reveals mRNA turnover strategies. Science, 367(6482): 1151–1156, 2020.

The Role of type VI secretion systems in marine communities

Astrid Stubbusch

ETH & EAWAG (Swiss Federal Institute of Aquatic Science and Technology)

Microbial interactions shape the composition and function of microbial communities with major implications for ecosystem ecology and biogeochemical cycles. Diverse antagonistic interactions exist between bacteria and contribute to temporal and spatial variability in the bacterial composition of marine communities. However, most interactions are studied in model organisms and our understanding of their prevalence and ecological impact remains limited. To bridge these knowledge gaps, we study the antagonism dynamics of single cells within simple marine communities composed of antagonists and non-antagonists of genus Vibrio using microfluidics coupled to time-lapse microscopy. The antagonist strain harbors the type VI secretion system (T6SS), a bacteriophage-like machinery that can translocate toxic effector proteins into neighbouring cells and is widespread among Gram-negative bacteria. We observe a strong link between environmental nutrient composition and antagonistic behaviours, confirmed by RNA sequencing, and find preliminary evidence for the role of T6SS in nutrient cross-feeding. We aim to identify the T6SS in isolate genomes of Vibrio as well as metagenome-assembled genomes of global marine datasets, which will allow us to study genomic adaptations to the T6SS and expression and prevalence patterns in natural communities. These findings will enable a strong understanding of the ecological role of the T6SS in marine ecosystems and contribute to a quantitative understanding of antagonistic interactions in fluctuating environments.

Integrative ensemble modelling of cetuximab sensitivity in colorectal cancer PDXs

Umberto Perron

Fondazione Human Technopole

Patient-derived xenografts (PDXs) are human tumour fragments engrafted into mice for preclinical studies and therapeutic testing. PDXs offer clear advantages over cell line models and even organoids in terms of structural complexity, heterogeneity, and stromal interactions. We characterised a cohort of (230) colorectal cancer PDX models on a genomic, transcriptomic, and methylomic level, as well as through screening with Cetuximab and various combinations of chemotherapics. We then assessed PDX model quality, stability, and concordance with publicly available CRC cohorts. Finally, we trained, interpreted, and extensively validated an integrated ensemble predictor of Cetuximab sensitivity. Our study highlights how large, varied PDX collections can be used to train highly accurate (up to AUC ¿ .9) and fully interpretable models of drug response. Our PDX-trained models 1) better recapitulate patient-derived diagnostic signatures than comparable cell-line-trained models, 2) they can be robustly validated across independent PDX cohorts and 'omics inputs, and 3) they suggest novel diagnostic biomarker candidates.

Posters

Spatial and temporal structure of staphylococcal communities determines capability to interact via the agr quorum sensing system

Julian Baer

Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich, University of Zürich, Zürich, Switzerland

Staphylococcus spp. are medically relevant bacteria that can exist in a benign (colonizing) or virulent (invasive infection) phenotype. The accessory gene regulator (agr) quorum sensing system of staphylococcal species often regulates virulence and the transition between commensal and pathogenic state. The 7 types of agr systems in Staphylococcus aureus (4) and Staphylococcus epidermidis (3) differ by minor alterations in the amino acid sequence of the signal peptide (autoinducing peptide, AIP). The agr system is rather specific to its own AIP type which results generally in cross-inhibition between agr types. The single-cell resolution of quorum sensing system activity and agr related interactions of temporally and spatially structured communities containing multiple agr types is unknown. Utilizing single-cell microfluidics and fluorescent agr reporter plasmids, we investigate differences in agr cross-inhibition between scenarios reflecting infection or colonization.

1. In an environment with unidirectional convective flow, only the upstream community can influence the downstream one. We monitor this unidirectional interaction with batch cultures used to feed single cells entrapped in a microfluidic device.

2. Two strains may start colonization or infection simultaneously. To quantify agr activity of bacteria in proximity, we use microfluidic devices containing a permeable membrane to separate the two strains to allow for bidirectional chemical interaction without physical contact.

3. Finally, we mix the strains together and allow unrestricted bidirectional interaction in large microfluidic chambers.

We anticipate reduced cross-inhibition in bidirectional interaction scenarios and potential overgrowth of one strain in mixed cultures. This could indicate a reciprocal virulence suppression resulting in expressing a commensal biofilm lifestyle of multi-strain colonization and infection.

Model-based evaluation of treatment during exercise in type 1 diabetes

Julia Deichmann

ETH Zürich

Type 1 diabetes (T1D) is an autoimmune disease where the body is unable to produce the glucose-lowering hormone insulin. In order to maintain normal blood glucose levels, administration of exogenous insulin is required. In addition, regular physical activity is recommended for people with T1D. However, glucose demand changes drastically during exercise and insulin sensitivity is affected for several hours during recovery. These changes lead to an increased risk of acute and late-onset hypoglycemia, and adjustment of treatment is necessary. Existing guidelines recommend additional carbohydrate intake and reduction of insulin use, but accurate adjustment remains an open problem.
Our aim was to evaluate currently recommended treatment adjustments and identify potential shortcomings and factors driving hypoglycemia risk in an in-silico study. As opposed to clinical studies, this allowed us to do extensive simulations and systematically examine a wide range of scenarios. We used a model of the glucoregulatory system including exercise metabolism to perform full-day simulations with exercise sessions of varying timing, intensity and duration. We applied different sets of guidelines, and analyzed the resulting acute and late-onset hypoglycemia risk as well as time spent in the target blood glucose range.
Our simulations show that additional carbohydrate intake and insulin reductions as recommended in clinical guidelines lead to a successful decrease of the acute hypoglycemia risk during and immediately after exercise. However, late-onset hypoglycemia risk, in particular during the following night, remains elevated for many scenarios and appears to be driven by the increased insulin sensitivity. Additionally, it is affected by the interplay between different dynamics from meals, insulin injections and exercise and their timing with respect to each other. These results suggest that further adjustments of T1D treatment to exercise are required, and similar in-silico studies could aid the development of new guidelines and the generation of testable clinical hypotheses.

Bayesian network models of intracranial aneurysm risk factors

Matteo Delucchi

ZHAW, Zürich, Switzerland

Intracranial aneurysms (IA) are balloon shaped expansions of cerebral arteries which are often asymptomatic and unrecognized. However, sudden rupture of IA leads to aneurysmal subarachnoid hemorrhage (aSAH), a severe form of stroke with often poor functional outcome and high mortality. Better understanding of rupture risk factors to guide clinical decision support is required to improve treatment.
In order to construct an explainable model for medical decision support, we learned two Bayesian network (BN) models for clinical evidence structuring. We found novel interdependencies among the risk factors and onto rupture, leading to new potential causalities. Discrete and additive BNs were compared to conventional clinical data analysis methods on a retrospectively collected, single-centric patient-level set (n=799) of the 9 major IA rupture risk factors.
Discrete BN models with heuristic structure learning and non-parametric bootstrapping resulted in a sparse BN. Mixed additive BN estimated with structural Markov chain Monte Carlo presented the same associations and some additional ones that highly improved clinical interpretability. Systematically restricting physically impossible associations favored the detection of a high number of causal relations supporting existing and proposing new theories of IA rupture risk factor interdependencies. For example, IAs in females are known to rupture at smaller size and later in life compared to males. Both BNs found this relation as indirect association with females showing less chance of hypertension and thus being more likely to develop individual IAs which tend to be smaller and less likely to rupture.
The present study proposes the feasibility of BN models to both quantify existing associations and to detect novel disease pathways in order to enable patient-specific decision support for IA treatment.

Cells sleep differently – The sleeping fruit fly at single cell resolution

Joana Dopp

VIB-KU Leuven Center for Brain and Disease Research

All animals, including Drosophila Melanogaster, sleep. Despite this fact, defining sleep remains challenging and is limited to imprecise measures, including behaviour and local field potentials. In order to truly understand what sleep is, we need to explore the underlying molecular changes that occur during sleep in distinct cell populations. To this end, we have generated a comprehensive single cell transcriptomic atlas of the adult central fly brain across the sleep/wake cycle. To eliminate technical variation of multiple batches with samples of different sleep or wake conditions, we applied a genetic multiplexing strategy. The dataset combining all runs comprises more than 106K single cells, of which 22% have been classified into one of 25 distinct cell types. Differential expression analysis, tree-based classification between sleep and wake conditions as well as template matching analysis revealed cluster-specific transcriptomic changes, suggesting that distinct cell populations sleep differently. Among all cell types, ensheathing glia show the most changed transcriptomic profile between sleep and wake conditions, while the transcriptome of several small cell populations correlates highly with an increase in sleep drive. Our finding of different transcriptomic signatures even within cell populations, highlights the need to study differences between sleep and wake at the single-cell rather than the bulk level. We have validated the differential expression results by using fluorescent in-situ hybridization and RNAi knockdown of selected candidate genes. We have also confirmed that the mRNA expression of key clock genes cycles between Zeitgeber times in clock neurons as well as perineurial and ensheathing glia, albeit with different phases. Here, we demonstrate for the first time that a transcriptomic signature of sleep is different for distinct cell populations in the adult fly brain. These findings bring us closer to defining sleep at a molecular level.

Sars-CoV-2 wastewater-based genomic epidemiology

David Dreifuss

CBG, BSSE, ETH Zürich

Wastewater-based epidemiology is gaining traction during the COVID-19 pandemic as a way to achieve unbiased, rapid and cost-effective community-level surveillance. With the emergence of spread of disruptive variants of concern, monitoring the genomics of the circulating pathogens is now of central interest, in addition to monitoring total case numbers. We demonstrate how routine sequencing of wastewater samples can achieve early detection, monitor the spread and even infer epidemiological characteristics of variants of concern. We go on to describe how this approach is used for epidemiological surveillance at the national level in Switzerland.

Read error correction for heterogeneous NGS samples using Dirichlet mixture models

Lara Fuhrmann

CBG, BSSE, ETH Zürich

RNA viruses have high mutation rates, short generation times, and large population sizes that result in diverse within-host viral populations, which affects disease progression and treatment outcomes. Therefore, it is important to characterize intra-host viral diversity. Next-generation sequencing (NGS) data allows assessing the genetic composition of viral populations. Dirichlet mixture models using Gibbs sampling for inference, as implemented in the tool ShoRAH, have proven to perform well compared to other methods in identifying low-frequency haplotypes and correcting sequencing errors. In order to converge to the target distribution the sampler needs to run long enough, as theory only guarantees convergence asymptotically. In practice, a high number of sampling iterations need to be performed, which can still result in the chain either not converging or running unnecessarily long. Using a Dirichlet mixture model to cluster reads around their (unobserved) haplotypes, we, therefore, developed two different inference approaches to efficiently learn the model parameters from the NGS data. We developed a Gibbs sampling approach using the effective sample size and the Gelman-Rubin diagnostic as

convergence measures to allow prior termination of the sampler if convergence criteria of the chain are met before the maximal number of sampling iterations is reached. Using variational inference, we developed a mean field approximation of the posterior distribution by assuming that the latent (unobserved) variables of the model are independent. In our simulation studies we find that the mean-field approximation outperforms Gibbs sampling in terms of runtime significantly. However, with enough sampling iterations, the Gibbs sampler achieves better recall and precision performance.

Reproducibility in genomics: Impact of bioinformatics Tools

Pelin Icer Baykal

CBG, BSSE, ETH Zürich

Reproducibility assessment is a crucial step to validate a discovery in all fields of science including in genomics and data-driven biomedical research. Reproducibility in genomics refer to the ability to produce consistent genomic result across replicates. Different factors and factor combinations can impact reproducibility in genomics and bioinformatics tools is the most important factor. Assessment of bioinformatics tools in terms of reproducibility requires a large-scale amount of technical replicates, which are reads obtained from different runs of sequencing. However, the limited amount of technical replicates challenges to assess bioinformatics tools. In this study we propose synthethic replicates which are computationally generated and are cost-free and efficient alternatives to technical replicates. We introduce two types of synthethic reads: i) randomly shuffled reads ii) reverse complemented reads. We assess read alignment tools and structural variant caller tools using these synthethic reads and compare the output to the ground truth (the output obtained from original data). Our results show that while some tools are not effected by input data variation some of other tools produce inconsistent results. We believe this method gives the opportunity to assess bioinformatics tools in terms reproducibility in a fast and efficient way.

Joint inference of repeated evolutionary trajectories and patterns of clonal exclusivity or co-occurrence from tumor mutation trees

Xiang Ge Luo

CBG, BSSE, ETH Zürich

Cancer progression is an evolutionary process shaped by both deterministic and stochastic forces. Multi-region and single-cell sequencing of tumors empower high-resolution reconstruction of the mutational history of each tumor. At the same time, it also highlights the extensive diversity across tumors and patients. Resolving the interactions among mutations and recovering the recurrent evolutionary processes may offer greater opportunities for successful therapeutic strategies. To this end, we present a novel probabilistic framework, called TreeMHN, for joint inference of repeated evolutionary trajectories and patterns of clonal exclusivity or co-occurrence from a cohort of intra-tumor phylogenetic trees. Through simulation studies, we show that TreeMHN outperforms existing alternatives that can only focus on one aspect of the task. By applying our method to an acute myeloid leukemia dataset, we find the most likely evolutionary trajectories and mutational patterns, consistent with and enriching known findings.

Marginalization in bayesian networks: exact and approximate inference

Fritz Bayer

CBG, BSSE, ETH Zürich

Bayesian Networks are probabilistic graphical models that can compactly represent dependencies among random variables and are widely employed in computational biology. Missing data and hidden variables require calculating the marginal probability of a subset of the variables, the so-called normalizing constant. While knowledge of the normalizing constant is crucial for various problems in machine learning and computational biology, its exact computation is generally not feasible for categorical variables due to the NP-hardness of this task. We develop a divide-and-conquer approach using the graphical properties of Bayesian networks to split the computation of the normalizing constant into sub-calculations of lower dimensionality, reducing the overall computational complexity.

Exploiting this property, we present an efficient and scalable algorithm for estimating the normalizing constant for categorical variables. The novel method is compared against state-of-the-art approximate inference methods in a benchmarking study, where it displays superior performance. As an immediate application, we demonstrate how marginalization can be used to classify incomplete data against Bayesian networks and use this approach for identifying the cancer subtype of kidney cancer patient samples.

Single-cell copy number calling and event history reconstruction

Jack Kuipers

CBG, BSSE, ETH Zürich

Copy number alterations are driving forces of tumour development and the emergence of intra-tumour heterogeneity. A comprehensive picture of these genomic aberrations is therefore essential for the development of personalised and precise cancer diagnostics and therapies. Single-cell sequencing offers the highest resolution for copy number profiling down to the level of individual cells. Recent high-throughput protocols allow for the processing of hundreds of cells through shallow whole-genome DNA sequencing. The resulting low read-depth data poses substantial statistical and computational challenges to the identification of copy number alterations. We developed SCICoNE, a statistical model and MCMC algorithm tailored to single-cell copy number profiling from shallow whole-genome DNA sequencing data. SCICoNE reconstructs the history of copy number events in the tumour and uses these evolutionary relationships to identify the copy number profiles of the individual cells. We show the accuracy of this approach in evaluations on simulated data and demonstrate its practicability in applications to two breast cancer samples from different sequencing protocols.

Unsupervised representation learning of regulatory motifs

Martin Grosshauer

Technische Universität München

With the advent of deep sequencing, genomic data has become exceedingly easy to obtain. Nevertheless, figuring out regulatory interactions relies on hand-crafted experiments, which are not as easily scalable. The aim of this project is to extend the recent advances in the field of natural language processing to the field of genomics while still retaining interpretability. The core of this project is learning the regulatory context of multi-species genomic data directly from sequence data in an unsupervised fashion. We used 10 million intergenic sequences derived from fungi genomes in order to train a neural network model that is able to pick up on the hidden regulatory motifs and the higher order logical combinations that arise therefrom. The long-term goal of this project is constructing an embedding for genetic sequences that is aware of regulatory motifs. This enriched representation would facilitate downstream tasks, since contextual information is explicitly embedded for the models as a feature to pick up on.

Modeling of Bladder Cancer Evolution from Field Effects

Marek Kimmel[1,3], Jolanta Bondaruk[2], Yujie Chen1, David Cogdell[2], Khanh Dinh[4], Roman Jaksik[3], Paweł Kuś[3], Sangkyou Lee[2], Ziqiao Wang[2], Peng Wei[2], Hui Yao[2], Li Zhang[5], Bogdan Czerniak[2]

[1]Rice University, Houston, TX, USA, [2]UT MD Anderson Cancer Center, Houston, TX, USA, [3]Silesian University of Technology, Gliwice, Poland, [4]Columbia University, New York, NY, USA, 5University of Cincinnati, Cincinnati, OH, USA

Bladder cancer is a common malignancy and an ideal model of carcinogenesis, facilitating studies of field effects. Simple anatomy of the bladder permits mapping of in situ preneoplastic conditions across the entire mucosa. Czerniak's laboratory developed an approach referred to as whole-organ histologic and genetic mapping that enables molecular profiling of cancer evolution from occult mucosal field effects to invasive disease. Whole bladder specimen is split into rectangular fields which undergo analyses such as

Whole Exome Sequencing Here we present a mathematical model of cancer mutational profile by a time-continuous Markov branching process. Parameters of cell progression across the transformation process carrying any given mutation on the top of previous mutations were estimated on the basis of variant allele frequencies in different mucosal areas. Modeling results and resulting estimates will be discussed. The estimates of mutation ages seem to correspond to the dormant and progressive phases of the bladder cancer.

Estimation of reference profiles for cell Type deconvolution

Zahra Nozari

University of Regensburg, Germany

Immune cell populations are usually small and their phenotype might be difficult to specify. In bulk gene expression, various cell entities are present, and contribute to the overall expression, but their underlying cellular composition is hard to quantify. Experimentally, it can be quantified using e.g. fluorescence activated cell sorting, cytometry by time-of-flight, and single-cell RNA sequencing. An alternative to experimental methods is bioinformatical approaches, namely digital tissue deconvolution. For digital tissue deconvolution, the set of cell types has to be specified beforehand, as reference profiles have to be designed. However, such reference profiles are commonly computed by getting the mean over of a set of single-cell profiles for each cell type. Different technologies for generating data such as microarray, bulk RNA-Seq and single-cell RNA-Seq have different protocols which degrades the performance of the deconvolution models. In this project, we overcome this problem by learning reference profiles directly from bulk data. We introduce a regularized least squares method, that finds reference profiles that optimize deconvolution performance, and simultaneously adapts for potential technology or tissue effects. For this we utilize bulk gene expression measurements, its cellular composition information and existing reference profiles.

The multilayer community structure of medulloblastoma

Iker Nuñez-Carpintero

Barcelona Supercomputing Center

Biomedical multilayer networks offer a wide range of possibilities for the interpretation of the molecular basis of diseases; a particularly challenging task in the case of rare diseases, where the number of cases is small in comparison with the size of the associated multi-omics datasets. In this work, we develop a dimensionality reduction methodology to identify the minimal set of genes that characterize disease subgroups based on their persistent association in the multilayer network at different levels of resolution.

We apply this approach to the study of a cohort of patients affected by medulloblastoma, a childhood brain tumor, using patient proteogenomic data. For that purpose, we constructed a multilayer gene graph that integrates general biomedical knowledge from 5 databases (Reactome, Recon3D Virtual Metabolic Human, BioGRID, KEGG BRITE "Target-based Classification of Compounds" and Monarch Disease Ontology) and performed a multilayer community trajectory analysis using the R package CmmD, that we implemented.

By applying CmmD, it is possible to consider co-existent community structures from different modularity resolution limits, as well as tracking different events throughout the associated process of network community decomposition. Such events can already be used as features for gene clustering or other machine learning tasks, such classification and prediction.

Our approach is able to recapitulate known medulloblastoma subtypes (accuracy ¿ 94%) and offers a clear characterization of the associated gene functions from each clinical subtype, with the downstream implications for diagnosis and therapeutic interventions.

We verified the general applicability of our method by applying it to an independent dataset, achieving very high performances (accuracy ¿ 98%). Overall, this approach opens the door to a new generation of multilayer-based methods able to overcome the specific dimensionality limitations of the rare disease datasets.

Extracting genetic interactions from CRISPR/Cas9 double-knock out modelling

Daniel O'Hanlon

EMBL-EBI

Double knock-out CRISPR/Cas9 screens are a promising tool in functional genomics for the identification of combination cancer drug targets, and to probe the mechanisms underlying genetic interactions.
Multiple sources of uncertainty, such as from imperfect knockout, off-target effects, and copy number variation, can complicate interpretation of results. This is particularly apparent in screens over many cell lines, where these effects can be context dependent. Furthermore, functional relationships between genes and their context can confound determination of universal genetic effects.
Here we propose a Bayesian model to extract robust genetic interactions. This probabilistically accounts for multiple sources of uncertainty and confounding variables using available control and calibration datasets. Additionally, interactions that are dependent on a specific context are controlled for via hierarchical priors.
The model is implemented in the flexible NumPyro framework, using stochastic variational-inference, and with the JAX/XLA backend to exploit many-core and GPU resources.

Cell type-specific gene co-expression modules define tumor heterogeneity in melanoma patients

Lars Bosshard & Michael Prummer

Nexus Personalized Health Technologies, ETH Zürich and Swiss Institute for Bioinformatics (SIB)

Gene co-expression networks are governing all cellular processes in health and disease. But the presence or absence of correlated gene pairs is difficult to interpret in bulk samples. For instance, the co-occurrence of two cell types can lead to an apparent co-expression of two genes even when they are completely independent within each individual cell. In single cell experiments, an observed correlation between a pair of gene is truly present within one cell. Here we use droplet-based single cell transcriptomics to discover disease-specific robust co-expression networks in different cell types from tissue biopsies of melanoma patients. We analyze each sample independently to arrive at patient-specific networks and subsequently compare them across the cohort. This way, we remove technical variability and perform what is called late integration of the data. To this end, co-expression sub-networks (aka, modules) are identified in each patient using community detection principles. Recurring as well as unique co-expression modules are compared to gene ontology terms to assign a biologically meaningful label. Any difference of the disease and cell type-specific module composition from common gene sets can provide new insight into disease causing mechanisms or novel treatment options. After all, many of the curated gene sets used for enrichment analysis were derived from bulk samples of healthy individuals or non-human model organisms or cultured cell lines. Moreover, patient-specific gene expression programs in various cell types may give rise to personalized treatment recommendations.

Interpretable neural network models for stroke patient outcome prediction

Beate Sick

ZHAW & UZH

Semi-structured data like a combination of images and tabular patient data like age or blood pressure are common in health care to predict a patient's outcome and estimate effect sizes of risk factors. Deep learning models have proven outstanding prediction performances on unstructured data but often lack interpretability in favor of prediction performance. Classical statistical models, on the other hand, provide interpretable effect estimates such as odds ratios but only apply to structured tabular data. Here, we present a novel class of models that join deep learning with classical statistical approaches and enable the integration of semi-structured data while achieving state-of-the-art results and yielding interpretable effect estimates. We demonstrate on data from stroke patients how to use this class of interpretable models for predicting the outcomes, identifying risk factors, and assessing the relative importance of the different data modalities for a high performance prediction.

Can proteome improve hidden breast cancer subtypes discovery with machine learning support?

Joanna Tobiasz

Silesian University of Technology

Breast cancer, a highly heterogenic disease, has five intrinsic molecular types determined in the early 2000s based on microarray gene expression patterns. Type identification is crucial for therapy choice, while molecular profiling allows for clinical outcome prediction. However, proteome, less prone to expression modifications than transcriptome, appears more reliable for cancer stratification. We applied various machine learning approaches on protein expression levels to detect breast cancer patient subpopulations, and we referred the results to molecular type established based on gene expression. Analyzed data come from TCGA-BRCA project. Protein levels were measured with the Reverse Phase Protein Arrays (RPPA) for 407 females with known intrinsic molecular cancer types. HDBSCAN

technique, graph-based Louvain community detection algorithm, and custom Divisive intelligent k-means algorithm (DiviK) were used to cluster the samples, and were combined with different dimensionality reduction methods, including Principal Components Analysis and Gaussian Mixture Model based filtration. The resulting patients' subpopulations were compared with type etiquettes provided by TCGA with PAM50 transcriptomics-based classifier. The molecular cluster integrity was assessed with the average $\eta 2$ effect size to quantify the grouping results. The higher value of the effect size, the higher the proteome profile similarity within clusters and higher dissimilarity among clusters is expected. Among all tested methods, the DiviK algorithm with GMM feature filtration applied to the distribution of the log2-scaled variance of protein abundance outperformed the remaining options. The high agreement of PAM50 and proteomics-based type assignments was observed. Our analyses revealed, independently of the classification method chosen, a high molecular diversity of Luminal A type. We identified three main Luminal A subtypes. Additionally, the subtype proteomics signatures were found and successfully tested in cross-validation scenario.

This study is supported by European Social Fund grant no. POWR.03.02.00-00-I029 [JT] and Silesian University of Technology grant for Support and Development of Research Potential [JP].

Exploring transcription variability along the cell cycle with metabolic labelling data and stochastic modelling

Dimitrio Volteras

Imperial College London

Single-cell transcriptomics technologies are central in studying stochasticity in heterogeneous cell populations. However, scRNA-seq data only reveal a static snapshot of single-cell gene expression profiles and do not directly capture dynamics of transcription. Recently developed metabolic labelling protocols in single-cell transcriptomics aim to monitor transcriptional dynamics by adding temporal resolution to scRNA-seq data. We are focusing on one of these protocols, named scEU-seq [1], that combines metabolic labelling of spliced and unspliced transcripts with cell-cycle ordering information. Analysis of scEU-seq data allows us to record mean changes and variability in gene expression levels along the cell cycle. We have developed a stochastic model of mRNA kinetics to be integrated with these data, that captures bursty synthesis, splicing and degradation of labelled and unlabelled transcripts, driven by cell-cycle dependent rate parameters. The model reveals that cell-cycle modulation of parameters such as burst size or burst frequency can explain the observed changes in expression of particular cell-cycle regulated genes. The objective of this study is to create a statistical inference framework to fit the model to the observed mean and noise trajectories on a genome-wide level. This framework can potentially elucidate how mechanistic parameters of gene expression are regulated by the cell cycle as well as provide links between stochasticity of bursty gene expression and cell-cycle variability.

Reference:
[1] Nico Battich, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S. Baron, Marvin E. Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. Science, 367(6482): 1151–1156, 2020.

The evolutionary trajectories of cancerised clones and their modes of selection

Xinyu Yang

University of Glasgow

Cancer is an evolutionary process characterised by profound intratumoral heterogeneity. Intratumoral heterogeneity can be quantified using in silco estimations of cancer cell fractions of tumour-specific somatic mutations. Here we propose a data-driven approach that extracts evolutionary signatures defining different modes of tumour evolution from individual tumour cancer cell fraction distributions. Based on results from 4146 TCGA samples, we demonstrate that these signatures identify tumours undergoing neutral evolution and uncover a dynamic switch between adaptive and innate immune processes as tumours escape immune surveillance. We also identify mutational processes underpinning different modes of tumour evolution and demonstrate that switching between adaptive and innate immune cell populations is accompanied by the clonal expansion of driver genes that modulate tumour-stroma interactions.

Machine Learning reveals patterns in transcriptomics and clinical datasets associated with atopic dermatitis in African children

Damir Zhakparov

Swiss Institute of Allergy and Asthma Research

Atopic Dermatitis (AD) is a chronic inflammatory skin condition that is often observed in children worldwide with higher prevalence and severity in African ancestry populations and is often associated with increased risk of other allergic disorders. AD is a complex disease that is believed to stem from the interplay of bacterial dysbiosis, dysregulated immune response, genetic factors, and environmental influence. Due to this heterogeneity, there are several disease endotypes phenotypes exist which makes the understanding of the underlying mechanisms and diagnostics problematic. Machine Learning (ML) is a powerful technique that has been shown to be successful in the field of allergy and skin disorders recently. It can identify hidden patterns in complex high-dimensional datasets as well as integrate different types of information to get insights. Data analysis using

ML can bring some light on how different features among immunological, genetical and epigenetical variables contribute to the development of AD in children and several disease phenotypes. Methods Our study cohort consisted of 150 African children that were stratified into groups based on their location (Urban, Rural) and diagnosis (Healthy, AD). The dataset consisted of bulk RNAseq results as well as questionnaire answers (clinical, environmental information). We performed ML Analysis utilizing classical approaches (Random Forest, Gradient Boosting, Support Vector Machines using sklearn Python library. Data integration was done with Canonical Correlation Analysis (CCA). Additionally, we performed standard statistical tests and differential gene expression analysis on these data. Results Using various feature selection methods, we have identified 18 genes that showed discriminative capacity among healthy and AD individuals. Moreover, such features as fuel type used for cooking, early age paracetamol exposure and family history of eczema in the questionnaire dataset were distinguishing controls from affected children. Finally, integration of both data types significantly increased the predictive ability of the model and showed genetical features that correlated with environmental exposure variables. Conclusion Our study provides insights into the pathogenesis of atopic dermatitis in children and into the variables that could be used as biomarkers in predicting susceptibility to AD.

Learning from the climate and its Nobel laureates

Peter Schuhmacher

Private Universität im Fürstentum Liechtenstenin

The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Physics 2021 "for groundbreaking contributions to our understanding of complex physical Systems"

Nobel laureate Manabe began his work with decidedly simple climate models. They came amazingly close to measurable observables, and above all, they helped to systematically develop systems thinking.

The same did Nobel laureate Hasselmann. But then he focused his work on piecing together the many components of the climate system: Atmosphere, oceans, terrestrial systems, cryosphere, three-diemsnional, time-dependent, and with feedback interactions.

The goal was to take a process-oriented view, and to minimize reliance on only statistical relationships. Interestingly, this huge global circulation models (GCM) with strictly physical view led to the similar explainer problem that is discussed today in comprehensive AI systems. Nobel laureate Hasselmann is therefore explicitly honored for providing the necessary framework for interpreting the results of complex climate models in the context of measured data.

Possible contribution to the workshop

The history of climate modeling has many similarities to biomedical research today. In consultation with the workshop organizers, aspects of this could be reviewed to provide a lateral view to the ongoing discussion.

(1) Noise: It is always clear to any climatologist that there is noise in everything he or she does that simply captures all aspects: Data (locations, sampling, representativeness, semantics,....), Models (model formulation, parameterization, implementation, numerics, spatial resolution, temporal resolution, error propagation, ...).

(2) Scaling: Scaling has emerged as an important methodological tool throughout fluid dynamics. This is less to be found in this clarity in other natural scientific fields. Scaling gives an insight in which temporal/spatial dimensions a phenomenon takes place, how it reacts to changes (external forcing), and how long it carries traces of this influence.

(3) The interplay of physical and statistical perspectives:The work of Nobel laureate Hasselmann shows that a sensible combination can lead the way. The complex GCM is the prior from a physical point of view from which the posterior must be judged against data. Also, the statistical model should not be arbitrary, but should have a physical reference.

Author

Peter Schuhmacher, today mainly involved with medical data in health care (semantics, real world data, real world evidence, efficiency, implementation) PhD in Climatology (ETH Zürich, Prof. Atsumu Ohmura) https://www.linkedin.com/in/peter-schuhmacher/

Participants

Constantin Ahlmann-Eltze EMBL Heidelberg
Michela Baccini Universita' di Firenze
Julian Baer Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich, University of Zürich, Zürich, Switzerland
Pablo Baeza Centurion German Cancer Research Center
Emmanuel Barillot Institut Curie
Fritz Bayer CBG, BSSE, ETH Zürich
Niko Beerenwinkel CBG, BSSE, ETH Zürich
Lars Bosshard Nexus Personalized Health Technologies, ETH Zürich and Swiss Institute for Bioinformatics (SIB)
Peter Bühlmann ETH Zürich
Joachim Buhmann ETH Zürich
Luciano Cascione Institute of Oncology Research /Universita' della Svizzera italiana
Blair Colyer City, University of London
Jérôme Dauvillier SIB - Swiss Institute of Bioinformatics
Julia Deichmann ETH Zürich
Matteo Delucchi ZHAW, Zürich, Switzerland
Andreas Deutsch TU Dresden
Joana Dopp VIB-KU Leuven Center for Brain and Disease Research
David Dreifuss CBG, BSSE, ETH Zürich
Francesco Ferraro Human Technopole
Pedro Ferreira CBG, BSSE, ETH Zürich
Nir Friedman Hebrew University of Jerusalem
Lara Fuhrmann CBG, BSSE, ETH Zürich
Jeroen Goedhart Amsterdam University Medical Centers
Angela Goncalves German Cancer Research Center (DKFZ)
Fridolin Gross ImmunoConcept, University of Bordeaux
Martin Grosshauer Technische Universität München
Haralampo Hatzikirou Khalifa University

Sven Hirsch ZHAW – Zürich University of Applied Sciences
Wolfgang Huber EMBL Heidelberg
Fern Hughes Imperial College London
Pelin Icer Baykal CBG, BSSE, ETH Zürich
Hans-Michael Kaltenbach ETH Zürich
Marek Kimmel Rice University
Jack Kuipers CBG, BSSE, ETH Zürich
Kieran Lamb University of Glasgow
Malte Londschien ETH Zürich
Xiang Ge Luo CBG, BSSE, ETH Zürich
Veselin Manojlovic City, University of London
Justyna Mika Silesian University of Technology
Khalimat Murtazalieva EMBL-EBI/University of Cambridge
Richard Neher University of Basel and SIB Swiss Institute of Bioinformatics
Zahra Nozari University of Regensburg, Germany
Iker Nuñez-Carpintero Barcelona Supercomputing Center
Daniel O'Hanlon EMBL-EBI
Sakshi Pahujani University of Cologne
Christian Panse FGCZ ETHZ—UZH
Umberto Perron Fondazione Human Technopole
Jonas Peters University of Copenhagen
Fernanda Pinheiro Human Technopole
Michael Prummer Nexus Personalized Health Technologies, ETH Zürich & Swiss Institute for Bioinformatics (SIB)
Francesco Puccioni Imperial College London
Srinithi Purushothaman University of Basel
Hélène Ruffieux MRC Biostatistics Unit, University of Cambridge
Peter Schuhmacher Private Universität im Fürstentum Liechtenstenin
Beate Sick ZHAW and UZH
Daniel Stekhoven ETH Zürich
Mats Stensrud The École polytechnique fédérale de Lausanne (EPFL)
Astrid Stubbusch ETH & EAWAG (Swiss Federal Institute of Aquatic Science and Technology)
Aleksandra Suwalska Silesian University of Technology
Eve Tasiudi Department of Biosystems Science and Engineering, ETH Zürich
Priyadarshini Thirunavukkarasu University of Basel
Joanna Tobiasz Silesian University of Technology
Alena van Bömmel Leibniz Institute on Aging – Fritz Lipmann Institute (FLI)

Susana Vinga INESC-ID, IST, ULisboa
Dimitrios Volteras Imperial College London
Aleksandra Walczak ENS Paris
Paula Weidemüller EMBL-EBI
Xinyu Yang University of Glasgow
Ke Yuan University of Glasgow
Hagen Zandt Vifor Pharma
Michael Zellinger ETH Seminar for Statistics, Peter Bühlmann Group
Damir Zhakparov Swiss Institute of Allergy and Asthma Research