Workshop on Computational Models in Biology and Medicine 2020

Joint workshop of the GMDS & IBS-DR working groups “Statistical Methods in Bioinformatics” and “Mathematical Models in Medicine and Biology”

February 4th – 5th, 2020

University of Bonn, Germany

Contents

Workshop outline

This workshop intends to bring together researchers from different research areas such as bioinformatics, biostatistics and systems biology, who are interested in modeling and analysis of biological systems or in the development of statistical methods with applications in biology and medicine.

Keynotes

- Fabien Crauste (Université de Bordeaux): “Mathematical Immunology: How to account for individual heterogeneity in theoretical models of the immune response?”
- Volker Roth (University of Basel): “Interpretable Machine Learning for Personalized Medicine”
- Dagmar Iber (ETH Zürich): “From Networks to Function – Computational Models of Organogenesis”
- Malte Lücken (Helmholtz Zentrum München): “Making the most of your data: Building a single-cell RNA-seq pipeline”

Workshop venue

The workshop takes place in the Lecture Hall Center of the University of Bonn in the Endenicher Allee 19 C, 53115 Bonn.

Organization

The workshop is jointly organized by the GMDS/IBS working groups “Statistical Methods in Bioinformatics” (speakers: Michael Altenbuchinger, Harvard T.H. Chan School of Public Health; Klaus Jung, University of Veterinay Medicine Hannover) and “Mathematical Models in Medicine and Biology” (speakers: Markus Scholz, University of Leipzig; Ingmar Glauche, Technische Universität Dresden), as well as Jan Hasenauer (University of Bonn) who is the local organizer.

Contact and local organization

Dr. Michael Altenbuchinger,
AG Statistical Bioinformatics,
E-mail: michael.altenbuchinger@ukr.de

Prof. Dr. Jan Hasenauer,
Interdiciplinary Research Unit Mathematics and Life Sciences,
University of Bonn,
E-mail: jan.hasenauer@uni-bonn.de

Support

The workshop is funded by the “Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)” and the “Deutsche Region der

Internationalen Biometrischen Gesellschaft (IBS-DR)" as well as the University of Bonn.

1 Program

Tuesday, February 4, 2020

11:30 Registration opens (with small lunch)

12:45–13:00 Welcome

Session 1: Dynamical modeling

13:00–13:40 Keynote lecture: Fabien Crauste
Mathematical Immunology: How to account for individual heterogeneity in theoretical models of the immune response?

13:40–14:00 Congxin Li
Non-equilibrium Transcriptional Regulation: Theory and Experiment

14:00–14:20 Marco Blickensdorf
Agent-based modelling of the impact of Pores of Kohn on infection dynamics of A. fumigatus in human alveoli

14:20–14:40 Yuri Kheifetz
Shiny platform for the analysis of complex individualized biomathematical models of hematopoiesis as well as for their implementation in controlling hematopoietical side effects during guided dose adaptation in chemotherapy

14:40–15:00 Thomas Zerjatke
Divisional behaviour of haematopoietic stem cells revisited: a quantitative comparison of label dilution techniques

15:00–16:00 Coffee break & poster session I

Session 2: Statistical modeling, machine learning and open topics

16:00–16:40 Keynote lecture: Volker Roth
Interpretable Machine Learning for Personalized Medicine

16:40–17:00 Michael Huttner
Diagnosis of Lymphoma subtypes using deep neural nets

17:00–17:20 Thomas Linden
Predicting comorbidities of epilepsy patients using big data from Electronic Health Records combined with biomedical knowledge

17:20–17:40 Helena Zacharias
A multi-source data integration approach reveals novel associations between metabolites and renal outcomes in the German Chronic Kidney Disease study

17:40–18:00 Johannes Hertl
Translating microbiome abundance patterns into patterns of metabolic function by integrating constraint based modelling with population statistics

18:30 Conference dinner in the Restaurant zur Bühne Bonn
Kapuzinerstr. 13, 53111 Bonn

Wednesday, February 5, 2020

Session 3: Genomics, proteomics and imaging

8:30–9:10 Keynote lecture: Dagmar Iber
From Networks to Function – Computational Models of Organogenesis

9:10–9:30 Moritz Kohls
Resampling metagenomics findings to assess reproducibility

9:30–9:50 John Wiedenhoeft
Dynamically compressed Bayesian Hidden Markov models using Haar wavelets

9:50–10:10 Rudolf Schill
Modelling cancer progression using Mutual Hazard Networks

10:10–11:10 Coffee break & poster session II

Session 4: Single-cell data analysis

11:10–11:50 Keynote lecture: Malte Lücken
Making the most of your data: Building a single-cell RNA-seq pipeline

11:50–12:10 Vladislava Milchevskaya
Distance metrics for single cell data

12:10–12:30 Martin Treppner
Generating Synthetic Single-Cell RNA-Sequencing Data from Small Pilot Studies using Deep Learning

12:30–12:50 Marian Schön
DTD: an R package for Digital Tissue Deconvolution

12:50–13:00 Closing remarks & poster award

13:00 Small lunch

2 Keynotes

Keynote 1

Mathematical Immunology: How to account for individual heterogeneity in theoretical models of the immune response?

Fabien Crauste

Université de Bordeaux

Immune cells allow a priori fast and efficient responses against non-self agents. They rely upon the ability of the organism to identify threats and trigger the most appropriate reactions. Cytotoxic immune responses aim in particular at inducing infected cell death, and to do so they integrate early on information about the nature of the infection in order to perform an appropriate differentiation program. This leads to an important inter-individual variability in terms of cell counts and temporal dynamics among individuals of a given population (for instance, mice or humans). Most theoretical models of immune responses, either mathematical or computational models, usually consider only population-aggregated values such as mean and standard deviation. I will discuss modeling approaches of specific immune responses, their ability to properly describe the differentiation process leading to the clearance of an infection, and how they can account for inter-individual variability without over-complexifying models.

Keynote 2

Interpretable Machine Learning for Personalized Medicine

Volker Roth

Universität Basel

Studying the effect of a therapeutic intervention on a patient is one of the major goals in computational medicine. In this talk I will approach this problem in the context of three related themes namely causal inference, decision making and interpretability. On the methods side, a specific focus will be put on information theoretic deep learning models for identifying a suitable representation of confounding in order to quantify treatment effects, and on tree-structured regularization methods that can be viewed as a means of gaining interpretability for decision-making. Further, I will present some applications for predicting medical outcomes of hospitalized septic patients and for predicting HIV therapy outcomes.

Keynote 3

From Networks to Function – Computational Models of Organogenesis

Dagmar Iber

ETH Zürich

One of the major challenges in biology concerns the integration of data across length and time scales into a consistent framework: how do macroscopic properties and functionalities arise from the molecular regulatory networks and how do they evolve? Morphogenesis provides an excellent model system to study how simple molecular networks robustly control complex pattern forming processes and how mechanical constraints shape organs. In my talk, I will focus on self-organizing principles in organogenesis, with a particular focus on lung and kidney development, as well as on epithelial organisation.

Keynote 4

Making the most of your data: Building a single-cell RNA-seq pipeline

Malte Lücken

Helmholtz Zentrum München

Single-cell RNA-seq has enabled gene expression to be studied at an unprecedented resolution. The promise of this technology is attracting a growing user base for single-cell analysis methods. As more analysis tools are becoming available, it is becoming increasingly difficult to navigate this landscape and produce an up-to-date workflow to analyse one's data. In this talk, I introduce the steps of a typical single-cell RNA-seq analysis, including pre-processing (quality control, normalization, data correction, feature selection, and dimensionality reduction) and cell- and gene-level downstream analysis. We will explore some tool choices for these steps and elaborate how tool choice can affect the biological interpretation of transcriptomic data. Finally, we will go over the current best-practices for single-cell RNA-seq analysis based on independent comparison studies that we formulated in our recent molecular systems biology paper, and introduce our best-practices analysis pipeline that is available at `https://www.github.com/theislab/single-cell-tutorial`. This talk is intended to serve as a workflow tutorial for new entrants into the field, and help established users update their analysis pipelines.

3 Oral presentations

3.1 Non-equilibrium Transcriptional Regulation: Theory and Experiment

Congxin Li

German Cancer Research Center (DKFZ)

Transcription is a dynamic non-equilibrium process. Here, we provide a theoretical framework of gene regulation based on quantitating how transcriptional factors (TFs) modulate the kinetic cycling of key gene states – transcriptionally inactive, active and refractory. Our theory shows that the sensitivity of gene response to a TF is controlled by the effect of the TF on the speed of gene-state cycle. A TF accelerating (slowing down) the gene-state cycle causes more (less) sensitive transcriptional response than the standard equilibrium model of gene regulation would predict. As a consequence, transcriptional activators that modulate the frequency of transcription bursts can trigger transcription at maximal rate through weak occupancy of their binding sites. We verify this prediction experimentally, using light-controlled gene activation by the GATA-type TF White Collar Complex (WCC) in Neurospora. Moreover, we find that burst frequency modulation by WCC allows for differential activation of its target genes independent of TF affinity. Data-based modeling indicates that differential gene regulation is due to the different gene activation rates after WCC binding, and we support this prediction using synthetic gene constructs. Finally, we show that refractory genes can be switched on more rapidly than non-refractory genes. In sum, our work demonstrates the relevance of a kinetic, non-equilibrium framework for understanding transcriptional regulation.

3.2 Agent-based modelling of the impact of Pores of Kohn on infection dynamics of A. fumigatus in human alveoli

Marco Blickensdorf, Sandra Timme, Marc Thilo Figge

Leibniz-Institut für Naturstoff-Forschung und Infektionsbiologie e. V. - Hans-Knöll-Institut (HKI)

The concept of systems biology constitutes a powerful tool to investigate biological systems. Thereby, wet-lab and dry-lab studies mutually support and complement each other. However, systems biology of infection often faces several problems on both sides of the systems biology cycle: First, since experiments can only be conducted in animal models the transferability of results to the human system is difficult. Second, even in animal experiments infection dynamics cannot be captured as whole such as the lung. However, virtual infection modeling provides the possibility to overcome the aforementioned limitations by integration of all available experimental data and thereby drives the research in systems biology of infection.

In the recent years we have developed a virtual infection model to investigate Aspergillus fumigatus lung infections. A. fumigatus is an environmental wide spread fungus that is opportunistic to humans and can cause severe infections in immunocompromised patients. Its spores, also called conidia, may reach the lower respiratory tract of the lung and, if not efficiently attacked by the immune system, cause invasive pulmonary aspergillosis with high mortality rates of 30%-90% making it a relevant target for research. Due to its complex interactions with the host immune system and its ability to adopt different morphologies many levels of pathogenicity have to be considered for development of effective therapy. For this purpose we could show the importance of alveolar signaling to the infection clearance and determine factors of their efficiency.

In our current study we investigate the role of so called Pores of Kohn (PoK) in the human alveolus. These PoK are connections of neighbouring alveoli in the lung and represent a possible gate for the entry or exit of immune cells such as alveolar macrophages. Despite these PoK are known for long time, their role in the lung for immunological signaling, air regulation and cellular communication is unknown and partly contradicting theories remain to be proven. To investigate the impact of different roles of PoK on various important dynamics in the infection clearance of A. fumigatus we use the previously developed virtual infection model. This model simulates cellular dynamics using an agent-based approach, which models the alveolus, the alveolar macrophages and the fungal spore in a realistic to-scale representation. Furthermore, molecular dynamics are modeled by PDE for chemokine secretion and ODE for intracellular receptor-ligand binding. To investigate several hypotheses regarding the functions of PoK we compared the infection dynamics of multiple model settings with different roles of PoK.

We were able to show how PoK impact on the dynamics of alveolar macrophages in terms of spatial distribution, migration dynamics and infection clearance. Furthermore, we quantified the impact of chemotacitic signaling pro-

cesses on the infection outcome and their dependency on various PoK settings and were able to put a new perspective on PoK.

3.3 Shiny platform for the analysis of complex individualized biomathematical models of hematopoiesis as well as for their implementation in controlling hematopoietical side effects during guided dose adaptation in chemotherapy

Yuri Kheifetz, Markus Scholz

Institut für Medizinische Informatik, Statistik und Epidemiologie, Leipzig University

Objectives:
Neutropenia, thrombocytopenia anemia are major side-effect of cytotoxic cancer therapies. The aim of precision medicine is to develop individual therapy modifications accounting for the individual's risk. Treatment modifications include chemotherapy dose adjustments, therapy postponement as well as supportive treatments. The development of individual therapy adaptations is a non-trivial task since hematological risks depend on many therapy-associated and individual factors. To solve this task, biomathematical mechanistic ordinary differential equations (ODE) models of thrombopoiesis, granulopoiesis, erythropoiesis and iron metabolism under various treatment scenarios have been developed during last decades by our research group in IMISE. Our models contain hundreds parameters, which can be properly estimated only by combination of data from clinical and various biological studies. Complexity of our model raised necessity in developing a novel concept of Graphical user interface (GUI) platform for scientific as well as for medical aims.

Methods:
Since maturations of different blood cells lines are interdependent and influenced from stem-cells-niches supporting osteoblasts, we combined a revised biomathematical model of granulopoiesis with our novel model of thrombopoiesis as well as with a model of osteoblasts/osteoclasts dynamics of other group and our novel model of lymphopoiesis. We implemented this combined model as well as our recent combined model of erythropoiesis with iron metabolism into two scientific and medical shiny tools. The scientific tool is intended for population and individual estimation of parameters, estimation of overfitting and parameters sensitivity. It proofs also models' stability at steady state. The tool enables direct combination of clinical studies' data with biological data based on the novel concept of virtual participation in other experiments, guarantying, that the resulting individual simulations would be consistent with current knowledge of different aspects of hematopoiesis from stem to circulating cells. Users can choose also mixed effects modelling option for parameters estimation. Regression modeling of models' parameters on biological covariates is implemented as well. The tool enables direct visualization of the individual and population time courses of blood cells during multicycle chemotherapy as well as under conditions of different biolog-

ical experiments. The medical tool is intended rather for clinicians. It enables also direct visualization of the individual time courses of blood cells during multicycle chemotherapy and supportive treatments through platelet transfusion, iron medication or growth factor applications such as thrombopoietin (TPO), erythropoietin (EPO), granulocyte colony-stimulating factor (G-CSF) or prednisolone. The users chose specific study and patients from an available database. The model-based prediction can be used to find minimal dose for the next cycle chemotherapy that control thrombocytopenia, neutropenia and anemia of prescribed degree. User can change and postpone the next-cycle dose manually and visualize respective predicted hematological side effects. The user can apply various supportive treatments as well. Long follow up period can be simulated, in order to predict dynamics of final restoration of hematological function after the treatment's end. The user approves and saves the finally selected protocol for the next cycle. It is possible either to work with each patient separately, or it is possible to make a simultaneous model-based prediction and nadir adjustment for a group of patients.

The tool possesses a framework to asses, improve and compare the performance of different models regarding prediction of next-cycle thrombocytopoenia at an individual level. We implemented also several semi-mechanistic models of other groups for the fair comparison of modeling and predictive power. Consequently, the user can simulate next-cycle scenarios with different models.

The tools are implemented in shiny, which works effectively on different operational systems. R-code calls C++ solvers of ODE in order to increase a computational speed.

Results&Conclusions:
The scientific tool, implemented originally in matlab enabled us to individualize and to combine our previous hematological models together. Due to the virtual participation algorithm, we fitted well on one hand individually time courses of thrombocytes and leukocytes from clinical data and on the other hand numerous independent biological studies, concentrated on all possible aspects of hematopoiesis and growth factors based feedbacks. New modeling framework made it much easier to simulate and visualize dynamics of different variables under different treatment scenarios, ensuring by this a quantitative understanding of hematopoiesis. We have validated the predictability potential of our medical tool by predicting nadirs of thrombocytes of 135 CHOP and CHOEP treated patients using thrombocytes dynamics from the previous cycles. We applied recently this framework to compare our thrombopoiesis model with other semi-mechanistic models. Our predictions were reliable and significantly more precise compared to those, done by implementing few existing popular simplistic model. Implementation of complex bio-mathematical models enabled description of new qualitative phenomena such as increasing toxicity of poly-chemotherapy in the late treatment cycles as well as the so-called first cycle effect. In the perspective, we hope that complex mechanistic models will be very useful in managing of pathway-specific anti-cancer drugs, targeting on

certain processes of precursors blood cells. On the other side, a parallel comparison with much simpler semi-mechanistic models enables to check critically the models' predictive potential.

3.4 Divisional behaviour of haematopoietic stem cells revisited: a quantitative comparison of label dilution techniques

Thomas Zerjatke

Institute for Medical Informatics and Biometry, TU Dresden

A fraction of haematopoietic stem cells (HSCs) rarely divides in vivo, with supposed cell cycle times being in the range of hundred days. In order to study this divisional behaviour of HSCs in mice, label dilution has emerged as a frequently used experimental approach: HSCs are provided with a certain amount of a labelling dye that is subsequently diluted upon cell division. Based on the resulting long-term dilution kinetics and the fraction of label-retaining cells, conclusions are drawn on the division frequency of the cells. There are mainly two common techniques for in vivo labelling without the need of cell transplantation, which thus allow to study unperturbed steady-state haematopoiesis: (i) DNA labelling by using thymidine analogues like BrdU that are temporarily provided to the animal by nutrition, or (ii) inducible histone labelling, where the expression of a fluorescently tagged histone is controlled by drug administration.

Here, we quantitatively compare the resulting kinetics of these two labelling approaches by applying ordinary differential equation models. We can show that two additional processes play a crucial role when using inducible histone labelling: leaky background label production, i.e. some expression of fluorescently tagged histone even after stopping the labelling period, and loss of fluorescence label due to protein degradation independent of cell division. These two processes qualitatively change the kinetics compared to a BrdU system and disregarding them can lead to severe misinterpretation in the analysis of label dilution data. We furthermore use our model simulations to show inherent limits of these labelling techniques in precisely estimating division times of slowly cycling HSC populations.

Our results demonstrate that an appropriate mathematical representation of the underlying biological processes is crucial in order to derive a meaningful interpretation of label dilution data.

3.5 Diagnosis of Lymphoma subtypes using deep neural nets

Michael Huttner

Institute of Functional Genomics, University of Regensburg

Oncologic precition medicine needs very exact sub-typing of tumors to enable the best possible therapy response. While pathologist are very good at this, diagnosing these tumors is a very laborious and analog process with limits in how many patients can be diagnosed, how much area of the slides can be looked at and how many subtypes are distinguishable by the human eye. But augmenting this process through digitalisation is possible, slides can be scanned digitally as images with about 10 gigapixels resolution.

We are developing a solution to automatically diagnose different subtypes of lymphoma or benign lymphadenitis based on these high resolution scans of lymph node slides. We use deep convolutional neural networks to train a “virtual pathologist” based on a data set of 628 scanned and labeled slides from 157 patients with 4 different immunohistochemistry stains each. Combined with a final random forest classifier our method is able to achieve 97% accuracy in patient diagnosis for our dataset.

3.6 Predicting comorbidities of epilepsy patients using big data from Electronic Health Records combined with biomedical knowledge

Thomas Linden

Bonn-Aachen International Center for Information Technology (B-IT); Fraunhofer SCAI Schloss Birlinghoven

Epilepsy is a complex brain disorder characterized by repetitive seizure events. Epilepsy patients often suffer from various and severe physical and psychological co-morbidities (e.g. Anxiety, Depression, Hypertension, etc.). While general comorbidity prevalences and incidences can be estimated from epidemiological data, such an approach does not take into account that actual patient specific risks can depend on various individual factors, including medication. This motivates to develop a machine learning approach for predicting risks of future comorbidities of epilepsy patients.

In this work we use Big Data from electronic health care records (around 100 Million raw observations), which provide a time resolved view on an individual's disease and medication history. We further enrich these data with information from several databases (DisGeNet, TTD, KEGG, DrugBank, SIDER etc.) to capture putative biological effects of observed diseases and prescribed medications and extracted more than 12,000 features.

We compared different machine learning approaches, such as Survival Random Forests and Deep Learning techniques for predicting future comorbidity occurrence after first epilepsy diagnosis. All candidate models were trained on around 20,000 patients. The deep learning model outperformed other methods. We also conducted an in-depth feature-importance analysis utilizing SHAP values and validated the model on an independent dataset of around 97,000 new patients.

Altogether we see this project as a step towards better personalized treatment of epilepsy patients.

3.7 A multi-source data integration approach reveals novel associations between metabolites and renal outcomes in the German Chronic Kidney Disease study

Helena Zacharias

University Medicine Greifswald, Department of Psychiatry and Psychotherapy

Omics data facilitate the gain of novel insights into the pathophysiology of diseases and, consequently, their diagnosis, treatment, and prevention. To this end, omics data are integrated with other data types, e.g., clinical, phenotypic, and demographic parameters of categorical or continuous nature. We exemplify this data integration issue for a chronic kidney disease (CKD) study, comprising complex clinical, demographic, and one-dimensional 1H nuclear magnetic resonance metabolic variables. Routine analysis screens for associations of single metabolic features with clinical parameters while accounting for confounders typically chosen by expert knowledge. This knowledge can be incomplete or unavailable. We introduce a framework for data integration that intrinsically adjusts for confounding variables. We give its mathematical and algorithmic foundation, provide a state-of-the-art implementation, and evaluate its performance by sanity checks and predictive performance assessment on independent test data. Particularly, we show that discovered associations remain significant after variable adjustment based on expert knowledge. In contrast, we illustrate that associations discovered in routine univariate screening approaches can be biased by incorrect or incomplete expert knowledge. Our data integration approach reveals important associations between CKD comorbidities and metabolites, including novel associations of the plasma metabolite trimethylamine-N-oxide with cardiac arrhythmia and infarction in CKD stage 3 patients.

3.8 Translating microbiome abundance patterns into patterns of metabolic function by integrating constraint based modelling with population statistics

Johannes Hertel

School of medicine, National University of Ireland, Galway

Introduction
The gut microbiome with its trillions of bacteria contributes crucially to human metabolism in health and disease by generating inaccessible nutrients, deactivating and activating drugs, and producing potentially harmful metabolites. Recent advances in sequencing techniques have given rise to a wealth of insights into patterns of gut microbiome composition. However, as species share metabolic capabilities and functions even across different phyla, it is unclear how changes in composition map onto changes in metabolic functions.

Methods
Herein, we apply community constraint-based reconstruction and analysis (COBRA) to map species abundance patterns onto patterns of metabolic functions. Community COBRA modelling solves optimization problems within the steady state solution space of the differential equation system, describing the metabolic capacities of the microbial community on the genome scale level. Its strengths of combining genomic data with condition specific constraints are specifically designed to deliver on the task of characterizing metabolic functions of microbial communities. However, COBRA modelling never was integrated systematically with population statistics approaches. The goal of this work is to set out the theoretical concepts allowing for the integration of COBRA modelling with population statistics and applying them to a recently published data set of a colon cancer case-control study (n=616), including fecal metabolomic and fecal metagenomic data (Yachida et al., 2019). Accordingly, we utilized metabolic reconstructions of hundreds of gut microbes (Magnúsdóttir et al., 2017) in combination with community modelling (Baldini et al., 2018) to predict metabolic outputs of microbial communities (Hertel et al., 2019), systematically relating patterns in predicted flux distributions to the multivariate structure of the fecal metabolome.

Results
Specifically, we show that species diversity translates into metabolic diversity, identifying key species for metabolic diversity whose contribute overly to metabolic diversity. Second, we demonstrate that microbial communities are generally unique in terms of metabolic functions contributing thereby to the individuality of human metabolism. Third, we reveal that the abundances of metabolite exchange reactions in communities are associated with faecal metabolite concentrations in a canonical way. Fourth, we demonstrate that pre-

dicted flux potentials through exchange reactions systematically explain variance in faecal metabolite concentrations. Finally, we shot that the correlation patterns among predicted fluxes reflected the patterns among the faecal metabolite concentrations.

Conclusions
Overall, we lay down the conceptual foundation for translating gut microbiome compositions into metabolic functions, applicable to large scale human cohorts. We demonstrate the validity of our framework in a proof-of-principle analysis integrating COBRA modelling with fecal metabolomics data on a freely available dataset of a case-control colorectal cancer cohort.

References

- Baldini F, Heinken A, Heirendt L, Magnusdottir S, Fleming RMT, Thiele I. 2018. The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. Bioinformatics; 35(13):2332-2334.
- Hertel J, Harms AC, Heinken A, Baldini F, Thinnes CC, Glaab E, Vasco D, Trenkwalder C, Krüger R, Hankemeier T et al. 2019. Integrated Analyses of Microbiome and Longitudinal Metabolome Data Reveal Microbial-Host Interactions on Sulfur Metabolism in Parkinson's Disease. Cell Reports; 29(7):1767-1777.
- Magnusdottir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jager C, Baginska J, Wilmes P et al. 2017. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol; 35(1): 81-89.
- Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K. et al. 2019. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Medicine; 25(6):968-976.

3.9 A bootstrap approach to estimate false positives in viral meta genomics

Moritz Kohls[1], *Ihsan Muchsin*[1], *Nicole Fischer*[2], *Paul Becher*[3], *Klaus Jung*[1]

[1]Institute for Animal Breeding and Genetics, University of Veterinary Medicine of Hannover
[2]Institute of Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf
[3]Institute of Virology, University of Veterinary Medicine Hannover

BACKGROUND:
Next-generation sequencing (NGS) is regularly used to identify viral sequences in a DNA sample of an infected host (Wang et al., 2013; Scheuch et al., 2015, Alawi et al., 2019). One part of most bioinformatics pipelines is to map sequencing reads or reads assembled to larger contigs to a database of known virus genomes and – if available – to the reference genome of the host. Due to similarities between viruses, mutations, mapping errors or conserved sequences the resulting lists of detected viruses can contain false positive and false negative findings. Very few approaches have been implemented to assess the error rates from bioinformatics virus detection pipelines. As one approach, Kruppa et al. (2018) have implemented a decoy database including false virus sequences. Additionally, this approach includes false sequences into the fastq-file from the sample. Here, the decoy strategy is refined to simulate more realistic reads.

METHODS:
The evaluation of reproducibility and assessment of error rates proceeds in four steps. The whole procedure is based on a sam-file resulting from mapping two paired fastq-files versus the set of all available virus reference genomes. The first step is then to estimate the statistical distributions from the sam-file. Next, these distributions are used to simulate two new paired fastq-files (step 2) which are then again mapped versus the viral reference genomes (step 3). From the new mapping results, mapping and error rates are calculated (step 4). There are four absolute frequencies counting all simulated reads belonging to a virus, reads that are mapped to any virus or in particular to a virus of the excel result list and reads that are mapped correctly. Additionally, there are six combinations of relative frequencies of mapping and error rates that are calculated by dividing the particular cardinal numbers of the previously mentioned four sets.

RESULTS & DISCUSSION:
In order to demonstrate the use of the new approach, the metagenomics pipeline is applied to the two samples files with known viral contents. First, the distributions of nucleobases, mapping qualities, read lengths, start positions and phred scores of the quality values are compared between the original sam-file and the sam-file which is obtained by mapping the artificially generated fastq-

files to the reference genomes. After that, the mapping rates and error rates like the false discovery rate are calculated and added to the excel result file. The latter is used as a measure for the quality of the decoy database to validate the results of the metagenomics virus detection pipeline. The evaluation of the new approach shows that error rates are helpful to judge the viral content of a sample. The approach is not restricted to our mapping pipeline but can be integrated into other virus detection pipelines, too.

References:

- Alawi, M., Burkhardt, L., Indenbirken, D., Reumann, K., Christopeit, M., Kröger, N., Lütgehetmann, M., Aepfelbacher, M., Fischer, N. & Grundhoff, A. (2019). DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. Scientific Reports, 9(1), 1-17.
- Kruppa, J., Jo, W. K., van der Vries, E., Ludlow, M., Osterhaus, A., Baumgaertner, W., & Jung, K. (2018). Virus detection in high-throughput sequencing data without a reference genome of the host. Infection, Genetics and Evolution, 66, 180-187.
- Scheuch, M., Höper, D., & Beer, M. (2015). RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC bioinformatics, 16(1), 69.
- Wang, Q., Jia, P., & Zhao, Z. (2013). VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PloS one, 8(5), e64465.

3.10 Dynamically compressed Bayesian Hidden Markov models using Haar wavelets

John Wiedenhoeft

University Medical Center Göttingen

Hidden Markov Models provide one of the standard approaches in the detection of copy-number variants (CNV). While frequentist methods such as Baum-Welch and Viterbi path are used extensively, Bayesian inference techniques to integrate over the entire space of latent variables have gained a reputation of being infeasible on genome-sized data such as WGS. In this work, we integrate Forward-Backward Gibbs sampling with a dynamic compression scheme based on Haar wavelet regression, in order to concentrate computational efforts on regions of more pronounced changes in the marginal state distributions. This leads to vastly improved memory requirements, speed and convergence of segmentation, and allows for the detection of CNV candidates within minutes.

3.11 Modelling cancer progression using Mutual Hazard Networks

Rudolf Schill

University of Regensburg, Department of Statistical Bioinformatics

Motivation
Cancer progresses by accumulating genomic events, such as mutations and copy number alterations, whose chronological order is key to understanding the disease but difficult to observe. Instead, cancer progression models use co-occurrence patterns in cross-sectional data to infer functional dependencies between events and thereby uncover their most likely order of occurrence.

So far, increasingly general models have been proposed for this task: events that depend on each other in linear chains (Vogelstein, 1988), trees (Desper, 1999), directed acyclic graphs (Beerenwinkel, 2007), groups (Raphael 2015; Cristea 2017) and cyclic networks (Hjelm, 2006). The latter were limited to facilitating dependencies and could not account for phenomena such as mutual exclusivity.

Results
We propose Mutual Hazard Networks (MHN) which consist of interlinked Cox Proportional Hazard models and may contain cycles. MHNs model events by their spontaneous rate of occurrence and by multiplicative effects they exert on the rate of successive events. These effects can be greater or less than one, i.e. facilitating or inhibiting. We provide an efficient inference algorithm that can learn MHNs from cross-sectional data alone.

MHNs compared favourably to acyclic models in cross-validated model fit on several datasets tested. In application to the glioblastoma dataset from The Cancer Genome Atlas, MHNs proposed a novel interaction in line with consecutive biopsies: IDH1 mutations are early events that promote subsequent fixation of TP53 mutations.

3.12 Distance metrics for single cell data

Vladislava Milchevskaya[1], *Nikolaos Papadopoulos*[2], *Johannes Soeding*[2], *Achim Tresch*[1]

[1] Institute of Medical Statistics and Computational Biology, University of Cologne
[2] Max-Planck Institute for Biophysical Chemistry, Goettingen

Many essential algorithms for single cell analysis such as clustering and trajectory reconstruction rely on distances defined on the single cell expression profiles. We show that standard distances such as Euclidean distance are heavily biased, especially in samples with low read coverage. To overcome this, we propose a new distance measure that is better calibrated and suited to detect biologically relevant changes in gene expression. Using a Bayesian approach, we define a probability for two cells to belong to the same cluster. In particular, we explicitly model what relevant change in gene expression is. This facilitates the incorporation of prior knowledge in an adaptive and interpretable way.

We evaluate the performance of our distance metric on simulated data in clustering and trajectory reconstruction tasks. Moreover, we apply it to scRNA-seq data obtained from mouse embryonic brain tissue (using known cell type markers as ground truth).

3.13 Generating Synthetic Single-Cell RNA-Sequencing Data from Small Pilot Studies using Deep Learning

Martin Treppner

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg

Background:
When researchers design experiments, they commonly start with a small pilot study to infer the properties of a full-scale investigation. In particular, they conduct simulations based on pilot data to determine the necessary sample size of future studies. Biologists increasingly use deep learning models for working with single-cell RNA-sequencing data. These models can learn a low-dimensional representation of expression patterns within cells—often interpreted as cell types. Here, we examine the ability of these methods to aid the design of experiments. More precisely, the models learn the structure of data from a pilot study and subsequently generate expression patterns for full-scale investigations.

Method:
For this task, we investigate two deep generative models. The first one – single-cell variational inference (scVI) – is a frequently used method in the field of scRNA-seq data analysis. Currently, expression patterns generated by scVI are sampled from the learned posterior distribution. This distribution, by definition, depends on the input data. Hence, staying too close to the input can potentially introduce bias when inferring to a larger population of cells. Consequently, we additionally suggest using scVI to sample from the prior distribution. This sampling strategy ensures that samples come from a diverse region of the input space. Moreover, we propose single-cell deep Boltzmann machines (scDBM) whose theoretical properties make it especially suitable for small data sets. Next, we take 30 subsamples of 500, 1000, and 2000 cells from two scRNA-seq data sets. Namely, the publicly available 10x Genomics data containing peripheral blood mononuclear cells (Zheng et al., 2017) and a currently unpublished CEL-seq2 data set from the hippocampus of three embryonic (E16.5) mice. Afterward, we train the models on these subsamples and generate synthetic expression patterns in the size of the original studies. Then, we apply UMAP for dimensionality reduction and Seurat clustering for the detection of putative cell types. We use the Davies-Bouldin Index (DBI) to evaluate the clustering performance of the original and synthetic data, respectively. Since scRNA-seq data is known to be extremely heterogeneous, we also compare the relative frequencies of generated cells per cluster. Lastly, we examine whether the synthetic data resembles univariate and bivariate structures of the original data on preselected marker genes.

Results:
We found that for clustering, scVI_posterior exhibits high variability when inferring from a small pilot study to a larger amount of cells. As expected, the variability decreases with increasing sample size. Expression patterns generated from scVI_prior and scDBM perform better in clustering tasks, as indicated by a lower DBI. Besides this, the models show mixed results when resembling the heterogeneity present in scRNA-seq data. In some scenarios, highly abundant cell types were overestimated, whereas less abundant cell types were not detected at all. Furthermore, we observe that all models properly learn the univariate distribution of marker genes, but have difficulties with capturing complex correlations between genes.

Conclusion:
We conclude that for making inference from a small-scale study to a large-scale experiment, it is advantageous to use scVI_prior or scDBM since the commonly used scVI_posterior produces expression patterns that are too close to the input data. Also, scDBMs show an additional advantage for small data sets – potentially due to its reduced complexity. All in all, we show that deep learning models can improve experimental design and therefore advance the replicability of scRNA-seq experiments. This improvement might amend translation to medical applications.

3.14 DTD: an R package for Digital Tissue Deconvolution

Marian Schön

Institute of Functional Genomics, Statistical Bioinformatics, University of Regensburg

Digital tissue deconvolution (DTD) estimates the cellular composition of a tissue from its bulk gene-expression profile. For this, DTD approximates the bulk as a mixture of cell-specific expression profiles.

Different tissues have different cellular compositions, with cells in different activation states, and embedded in different environments. Consequently, DTD can profit from tailoring the deconvolution model to a specific tissue context. Loss-function learning adapts DTD to a specific tissue context, such as the deconvolution of blood, or a specific type of tumor tissue (Görtler et al, 2018). With our R package DTD, we provide a framework for optimizing deconvolution models to a tissue context. The optimization is done on artificial mixtures with known cellular compositions. Artificial mixtures are generated using labelled scRNA-Seq data.

We provide software for model learning, for its validation and visualization, and for applying the DTD models to new data. In this talk, we demonstrate a DTD analysis step-by-step. We show how DTD optimizes a deconvolution model to compensate for

(i) missing reference profiles,

(ii) small cell fractions, that are hard to quantify,

(iii) cell types whose expression profiles are highly correlated.

Additionally, we show how DTD models deconvolute the cellular composition of bulk expression data. DTD is available under
`https://github.com/MarianSchoen/DTD`.

4 Poster presentations: session I

4.1 Bond graph semantics – biophysically and thermodynamically consistent modularisation of physiology

Niloofar Shahidi

University of Auckland (Bioengineering Institute), New Zealand

The major focus of my study is to explore and develop standards, tools and databases that enable the reuse of reproducible mathematical models of physiological processes. As the modelling of physiological and biological systems has spread, grown and become more comprehensive, a need for reusability, understanding, composing and decomposing of the existing models has emerged. Considering the growing complexity of such models which are implemented in a variety of formats (e.g., SBML, CellML, C, R, Matlab), endeavours to archive and share such models as reusable modules for composition into larger, more complex, systems have similarly evolved.

The goal of my project is to enable the automated composition of arbitrary models discovered in an existing repository, the Physiome Model Repository (PMR, https://models.physiomeproject.org). To achieve this will require novel model representation concepts to be developed, combining the mathematical and biological semantics with modern data science and informatics tools. To automate the model composition process, we are investigating the use of graph-based data structures to directly and simultaneously represent bond-graph-based implementations of the computational models and biological semantics, along with the derived computational simulation experiments. The use of energy-based bond-graph principles will, furthermore, guarantee that the resultant model will similarly obey thermodynamic and physical laws. Thus providing a level of confidence in automated model composition that is currently lacking.

The bond-graph approach is especially valuable when models are required to span multiple types of physics (electrics, mechanics, fluids, chemical, etc.) – almost by definition this makes it an invaluable tool in modelling physiological systems.

An adaption of bond-graph theory to the fields of systems biology and computational physiology has been developed as a Python library – BondGraphTools – by P. Cudmore, PJ Gawthrop, and EJ Crampin at the University of Melbourne (https://arxiv.org/abs/1906.10799). We are currently using BondGraphTools as our primary tool for creating modular biological models.

Following the initial scope of this study, “toy” models of electrical circuits, mechanics, and chemical systems along with the cardiac electrophysiology, circulatory and renal transport have been modelled and verified by using BondGraphTools. These are chosen to provide proof of concept of model composition across multiple physical systems in intuitively understandable systems and

in which published models exist which can be used to verify and test the automatically composed models. We are now looking to explore how these reusable bond-graph modules can be combined with semantic annotations to describe relevant biology and enable their integration into existing CellML-based model discovery and composition frameworks (e.g., https://doi.org/10.1186/s12859-019-2987-y and https://doi.org/10.1093/bioinformatics/bty829).

4.2 The prediction of adverse events in chronic kidney disease patients with LASSO Cox proportional hazard regression

Sahar Ghasemi

Clinic and Polyclinic for Psychiatry and Psychotherapy, University Medicine Greifswald

Chronic kidney disease (CKD) patients are at a high risk of experiencing major adverse events, including progression to end-stage kidney disease (ESKD), cardiovascular (CV) events, and death. Timely identification of CKD patients at risk of experiencing future adverse events is a prerequisite for the initiation of targeted treatments, thus lowering patient mortality and morbidity, as well as associated health care costs. The risk prediction for individual CKD patients could be facilitated by employing adverse event risk equations specifically optimized for the CKD setting.

We explore the potential of machine learning algorithms to develop novel risk equations for CKD patients. Our study cohort comprises 5,215 CKD patients enrolled into the German Chronic Kidney Disease (GCKD) study (mainly stage G3) , who have been prospectively followed-up for four years. Patient parameters assessed at baseline were used as possible predictors. End-points included initiation of dialysis, renal transplantation, or a major nonfatal or fatal CV event. To facilitate easy transfer into clinical practice, our set of possible predictors were restricted to clinical chemistry, demographic, and/or disease history parameters readily available from routine CKD patient examinations. The adverse event risk equations were developed employing the state-of-the-art Cox proportional-hazards least absolute shrinkage and selection operator (Cox PH LASSO) algorithm and were subsequently tested in a rigorous cross-validation approach.

Our machine-learning-based adverse event risk equations showed overall good predictive performances, assessed by concordance indices (c-indices) ranging from $0.71 - 0.90$, thereby even outperforming state-of-the-art risk equations. The proposed risk equations facilitate the timely identification of CKD patients at risk of experiencing a major adverse event, only relying on readily available patient parameters.

4.3 A statistical methodology for data-driven partitioning of infectious disease incidence into age-groups

Itai Dattner

University of Haifa, Israel

Understanding age-group dynamics of infectious diseases is a fundamental issue for both scientific study and policy making. Age-structure epidemic models were developed in order to study and improve our understanding of these dynamics. By fitting the models to incidence data of real outbreaks one can infer estimates of key epidemiological parameters. However, estimation of the transmission in an age-structured populations requires first to define the age-groups of interest. Misspecification in representing the heterogeneity in the age-dependent transmission rates can potentially lead to biased estimation of parameters. We develop the first statistical, data-driven methodology for deciding on the best partition of incidence data into age-groups. The method employs a top-down hierarchical partitioning algorithm, with a metric distance built for maximizing mathematical identifiability of the transmission matrix, and a stopping criteria based on significance testing. The methodology is tested using simulations showing good statistical properties. The methodology is then applied to influenza incidence data of 14 seasons in order to extract the significant age-group clusters in each season.

4.4 Identification of genes associated with the onset of Parkinson's disease from diverse age dependant gene expression datasets

Pankaj Dholaniya

University of Hyderabad, India

Parkinson's disease (PD) has second largest mortality rate in neurodegarative disorders occurring more frequently among elderly people. Aging is considered as one of the greatest risk factor for Parkinson's disease with onset of the disease at an average age of 60 years. In the present study we have analyzed seven different gene expression datasets of varying age intervals belonging to individuals with normal aging and with the symptoms of Parkinson's disease. The datasets were taken from GEO database and specifically selected for substantia nigra. Along with the identification of differentially expressed genes, we have also identified the significantly correlated gene pairs using network-based approach. The network analysis has led to the identification 38 putative genes, which could be associated with the onset of PD, of which most of genes were previously reported in PD and few novel genes were also identified.

4.5 Linear system identification from ensemble snapshot observa

Atte Aalto

University of Luxembourg, Luxembourg

Gene expression models enable prediction of the effects of perturbations on the system, discovery of disease mechanisms, targets for drugs, and so on. Single-cell experimental techniques have recently become more and more common. They enable gene expression measurements on single cell resolution for thousands of cells at a time. Unfortunately, the cells are destroyed in the measurement process, and so the data consist of snapshots of representative subpopulations, measured at different times. The sheer amount of data produced by single-cell techniques far exceeds what is obtained with older bulk techniques, but it is not clear how such data should be used in modeling. In this work, we study linear system identification from single-cell data. We introduce a method based on tracking the evolution of the distribution of cells over time. The idea is to look at two consequtive snapshot observations, propagate the earlier observations through the candidate model, and compare the propagated observation distribution to the later observation distribution using the discretized Jensen-Shannon divergence.

4.6 Dynamical Modelling of Single-Cell RNA-Sequencing Data by the Chemical Master Equation

Stefano Magni, Alexander Skupin, Jorge Goncalves

Luxembourg Centre for Systems Biomedicine - University of Luxembourg, Luxembourg

Single-cell RNA-Sequencing (sc RNA-Seq) is a family of transcriptomics experimental techniques which are revolutionizing the way we can study cells at the level of gene expression. Sc RNA-Seq techniques allow nowadays to measure gene expression of almost each gene in the genome, for thousands of individual cells simultaneously. The outputs of such measurements are so called digital gene expression matrices, high-dimensional data which contain a great amount of information. Moreover, this type of data has one more dimension than bulk RNA-Seq methods, i.e. that of individual cells within a cell population.

Since these experimental techniques have been developed only in the last decade, the statistical data analysis methods and the mathematical modelling approaches tailored to them are only beginning to be widely developed. Nevertheless, these methods play a crucial role in extracting the information contained in these data. In particular, mathematical modelling is essential to exploit these data to test our understanding of the underlying system and generate new hypothesis from them.

Since sc RNA-Seq techniques measure gene expression individually for a population of single cells, the data so generated can be regarded as distributions of cells across gene expression. Moreover, while on the one hand usually it is possible to generate sc RNA-Seq data for only a very limited number of time-points, on the other hand at every time-point a subset of the whole population of individual cells is measured, leading to a distribution of cells over gene expression. Such distributions measured at multiple time points contain information about the dynamics of the underlying biological process.

To model this dynamics, among the various approaches which can be inspired from physics here we employ the formalism of the chemical master equation. This implies to use a set of coupled ordinary differential equations to describe the time evolution of the probability distribution of gene expression. This approach provides thus a probabilistic description, which is particularly suitable for this process since at the microscopic level gene expression is an intrinsically stochastic phenomenon.

We thus assume a simple low-dimensional class of toy models describing gene expression. We then generate synthetic sc RNA-Seq data via the Gillespie algorithm, to have a test dataset of which we know the underlying system. We further extract snapshots from these trajectories, to represent typical time-points measurements in sc RNA-Seq data. The chemical master equation is then used to generate from the assumed model class the distributions of cells across gene expression, to be compared with these data. To infer the model parameters which best fit the synthetic data, we employ Bayesian inference.

We thus define the Likelihood function for the model parameters given the data. This function can be used to obtain the posterior probability distribution for the model parameter values. We do so by Markov Chain Monte Carlo methods, which we use to estimate the posterior. We then proceed to evaluate the accuracy of our inference approach in recovering the original parameter values from the synthetic data.

Overall, here we set up and illustrate with a simple model a framework which can be used to develop dynamical probabilistic models from single-cell RNA-Seq data. Such models can be useful for several applications. A potential application of this framework is to infer regulations between genes from sc RNA-Seq measurements (gene regulatory network inference). Furthermore, this framework could be applied to investigate the biological process of cell fate determination, i.e. how a population of cells evolve in time from one cell type, to one or more other cell-types.

4.7 Reproducible modular kidney model in CellML

Leyla Noroozbabaee

AUCKLAND BIOENGINEERING HOUSE, New Zealand

We are using A. Weinstein's nephron model as a base model. Even though the model includes many biological details and also is able to explicate various functions of the kidney; but, the reproducibility remains the biggest issue with the fore-said model. The provided information in Weinstein et al. publications doesn't guide the researcher through a track to create the same kind of mathematical model which is able to produce similar results. In the past decade several attempts have been made to create reproducible and reusable biological kidney models in order to let other researchers to build a more comprehensive model based on the previous works. Regarding this matter, we came to this conclusion that the above-mentioned kidney model doesn't have the required necessities as a reproducible and reusable model. As such, considering their achieved results, we have focused on creating a clear, reproducible, and reusable modular-based model with meaningful parameters and variables to make it straightforward for other researchers with different backgrounds to investigate the model and add further developments as needed.

4.8 Distribution-free differential expression analysis for scRNA-seq data across patient groups

Erika Dudkin

Faculty of Mathematics and Natural Sciences, University of Bonn

Single cell RNA-sequencing (scRNA-seq) data provide insights into gene expression profiles of individual cells on a large scale. This contributed in recent years substantially to the understanding and identification of cell types and differences between them. To unravel differences between cell populations, a multitude of differential expression (DE) methods has been introduced to compare clusters of cells. However, these methods are not suited for the identification of differences between patient groups for which scRNA-seq data are available. Typically, DE-analysis was performed on a single sample or across multiple samples, leading then to cross-condition analysis. The emergence of scRNA-seq datasets with replicated multi-conditions, for example multiple patients of one condition versus multiple patients of a second condition, demands the development of new particular methods which cover this issue.

In this work, we present a method for the statistical comparison of replicated multi-conditions. The method uses Wilcoxon rank sum test for the pairwise comparison of samples. Differences between patient combinations are evaluated while taking all single cell read counts into account. After calculating the test statistic, its significance is determined with a permutation test.

The method is applied on a scRNA-seq dataset with multiple controls and chronic obstructive pulmonary diseased (COPD) patients. Differentially expressed genes were identified and underlying cellular mechanistic hypothesis of COPD could be confirmed by performing gene set enrichment analysis on the resulting DE-gene list.

4.9 Investigating mechanisms of cell fate by mathematical modelling

Françoise Kemp

LCSB, Luxemburg

Cell fate is the process by which cells adapt to their environment by changing their gene regulatory network (GRN) and plays a central role in cell differentiation but also in disease development. Despite intensive research and advancements of experimental approaches such as diverse omics technologies during the last decades, our understanding of underlying principles is still rather limited because of the complex interactions of the plethora of biological entities. Mathematical modelling represents a useful tool to test mechanistic hypotheses and to identify underlying principles of complex biological dynamics. We develop a generic model for cell (trans)differentiation based on ordinary differential equations describing the changes of the GRN in dependence on external conditions such as growth factors or paracrine signalling. We apply our model to epithelial to mesenchymal transitions (EMT) and characterize how cooperativity between subpopulations result in stable population dynamics. Our framework allows to quantify the underlying epigenetic landscape and to characterize cell differentiation also in the context of critical transitions.

4.10 Automatic generation of priors for large-scale dynamic models

Jakob Vanhoefer

Universität Bonn

Background:
Mechanistic ordinary differential equation models are a powerful tool to understand biological networks. Recent advances in computational techniques facilitate the study of models with thousands of state variables and parameters, allowing for a more detailed and realistic description of the underlying process. However, when parametrizing a large-scale model, even large data sets do often not provide enough information to provide reliable estimates for all parameters. One way to tackle this problem is to use a Bayesian setting and include prior knowledge obtained from the literature or previous investigations. This facilitates the implicit integration of additional data into the own data analysis. However, the construction of priors for models with a large number of parameters can be tedious.

Method:
Here, we present a tool for the automatic construction of priors for SBML models annotated by terms from the Systems Biology Ontology. The tool identified for each parameter the reaction type and kinetic law of the reactions using the parameter. Thereby we infer the biological meaning of a parameter (e.g. protein decay rate) and assign a corresponding prior distribution. The prior information is extracted from data bases like BRENDA.

Results:
The resulting prior distributions are written in an PEtab parameter file. Thereby the priors can be used by any toolbox that can import problems specified in the PEtab format. In the future we plan to move from coarse grained to more fine grained and specific priors for the individual parameters.

4.11 No Noise No ABC

Yannik Schaelte

Helmholtz Center Munich, Institute of Computational Biology

Approximate Bayesian Computation (ABC) is an increasingly popular method for likelihood-free parameter inference in systems biology and other fields of research, being easy to use and broadly applicable to complex stochastic models. However, the approximation error made by this method is often not clear, in particular when there is measurement noise, which is rather common in biological applications.

In this contribution, we thus firstly raise awareness of the problem by illustrating how neglecting noise in ABC, which is unfortunately easy to do, can yield highly erroneous parameter estimates. Secondly, we discuss practical ways of correcting for it. In particular, we present a novel self-tuned sequential algorithm that allows to do exact likelihood-free inference for general noise models and is orders of magnitude more efficient than existing approaches on test problems. Thus, the proposed algorithm could improve the accuracy of parameter estimates for a broad spectrum of applications.

4.12 Learning the Topology of Latent Signaling Networks from High Dimensional Intervention Effects

Zahra Sadat Hajseyed Nasrollah

Institute of Medical Statistics and Computational Biology(IMSB), Faculty of Medicine University of Cologne

"Data based learning of the topology of molecular networks, e.g. via Dynamic Bayesian Networks (DBNs) has a long tradition in Bioinformatics. The majority of methods take gene expression as a proxy for protein expression in that context, which is principally problematic. Further, most methods rely on observational data, which complicates the aim of causal network reconstruction. Nested Effects Models (NEMs – Markowetz et al., 2005) have been proposed to overcome some of these issues by distinguishing between a latent (i.e. unobservable) signaling network structure and observable transcriptional downstream effects to model targeted interventions of the network. In this study we developed a more principled and flexible approach for learning the topology of a dynamical system that is only observable through transcriptional responses to combinatorial perturbations applied to the system. More specifically, we focus on the situation in which the latent dynamical system (i.e. signaling network) can be described as a network of state variables with nonlinear activation functions. We show how candidate networks can be scored efficiently in this case and how topology learning can be performed via Markov Chain Monte Carlo (MCMC). We extensively tested our approach by reconstruction of simple network motifs over a wide range of possible settings. Exertion of proposed nonlinear dynamics on the breast cancer proteomic dataset from the DREAM8 challenge shows cell line specific interactions between PIK3\AKT\mTOR and MAPK pathways . In other context, we aimed to expose interactions among 20 proteins that are involved in EGFR signaling and associated with drug sensitivity in Non Small Cell Lung Cancer (NSCLC). Results have been compared with reconstructed network from classical NEM model (Paurush et al., 2016)."

4.13 Estimation of properties of the heart from beat to beat pulse pressure variations in atrial fibrillation using a mechanistic mathematical model: particle methods from bench to bedside

Maximilian Oremek[1], *Andreas Hoeft*[2], *Sven Zenker*[1,3,4]

[1]AG Angewandte Mathematische Physiologie, Klinik und Poliklinik für Anästhesiologie und Operative Intensivmedizin, Universitätsklinikum Bonn
[2]Klinik und Poliklinik für Anästhesiologie und Operative Intensivmedizin, Universitätsklinikum Bonn
[3]Medizinisch-Wissenschaftliche Technologieentwicklung und -koordination (MWTek), Kaufmännische Direktion, Universitätsklinikum Bonn
[4]AG Angewandte Medizinische Informatik, Institut für Medizinische Biometrie, Informatik und Epidemiologie (IMBIE), Universität Bonn

Objectives:
In previous work, we were able to show that mechanistic mathematical models of physiology afford the possibility of quantifying unobservable but clinically relevant physiologic parameters from routine data (Kiefer et al., 2019). In that setting, we were unable to robustly and automatically quantify estimation uncertainty. This, however, is an important requirement when trying to bring these techniques to the bedside for clinical decision support. We therefore explored the usage of more complex Monte Carlo methods, applied to an extension of the model. The goal is to establish a tool chain which may allow the inference of properties of the heart from beat-to-beat pulse pressure variations in atrial fibrillation, whilst also allowing to quantify uncertainty of the parameters robustly for bedside usage.

Methods:
Using invasive arterial blood pressure and ECG data from the PhysioNet MGH/MF database (Welch et al., 1991), a time discrete dynamical model (Zenker et al., 2007) was evaluated in 3 sinus patients and 17 AF patients. Additionally, the same tool chain was used to extract data from an AF patient from the local patient database to demonstrate viability of workflows in a more translational setting. As stroke volumes are not routinely measured, pulse pressures (PP) were used as a surrogate marker. Data extraction used the AcuWave Software Suite (Begerau et al., 2017). To quantify uncertainty, the Population Monte Carlo approximate Bayesian Computation Algorithm (abcPMC) was used (Beaumont et al. 2009).

Results:
The AcuWave Software Suite allowed for visual verification of automated beat detection. A minimum of 202 consecutive beats could be extracted for further processing from all patient datasets. Implementation of the above mentioned abcPMC allowed automated inference of parameters and their distribution without manual tuning of the proposal distribution in all AF patients with reasonable efficiency and stability. The inferred diastolic stiffness, mitral valve resistance,

and filling pressure parameter modes were found to be within previously described physiological limits in AF patients.

Conclusion:
Unobserved properties of the heart could be estimated within the physiological range with the employed sampling method with improved tuning and automation properties compared to the previous approaches. Inferred posteriors have limited direct probabilistic interpretability, however, due to the L1 metric used in constructing the likelihood surrogate. Additionally, we illustrated that our visualization and analysis toolchain can be transferred seamlessly from research databases to real clinical data, paving the way for validation studies to confirm these findings.

References

- Beaumont, M. A., et al. “Adaptive Approximate Bayesian Computation.” Biometrika, vol. 96, no. 4, 2009, pp. 983–990., doi:10.1093/biomet/asp052.
- Begerau H, et al. The AcuWave Software Suite: a modular analysis and visualisation tool to facilitate the evaluation of derived parameters for researchers and clinicians in acute care. In: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, Hrsg. 62. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Oldenburg, 17.-21.09.2017. Düsseldorf: German Medical Science GMS Publishing House; 2017. DocAbstr. 254
- Kiefer, Nicholas, et al. “Model-Based Quantification of Left Ventricular Diastolic Function in Critically Ill Patients with Atrial Fibrillation from Routine Data: A Feasibility Study.” Computational and Mathematical Methods in Medicine, vol. 2019, 2019, pp. 1–11., doi:10.1155/2019/9682138.
- Welch JP, Ford PJ, Teplick RS, Rubsamen RM. The Massachusetts General Hospital-Marquette Foundation Hemodynamic and Electrocardiographic Database – Comprehensive collection of critical care waveforms. J Clinical Monitoring 7(1):96-97 (1991).
- Zenker, S., J. Rubin, and G. Clermont. 2007. “From Inverse Problems in Mathematical Physiology to Quantitative Differential Diagnoses.” PLoS Comput.Biol. 3 (11): e204.

5 Poster presentations: session II

5.1 Statistical comparison of two Alzheimer‘s disease cohorts and validation of an artificial intelligence model to predict disease diagnosis

Colin Birkenbihl

Fraunhofer SCAI

Artificial intelligence approaches pose a great opportunity for pre-symptomatic disease diagnosis. One option for building the appropriate models are observational cohort study data. However, recruitment criteria can bias statistical analysis of clinical cohort studies and impede model application beyond the training data. To evaluate if and how data from independent clinical cohort studies differ, we systematically compared ADNI and AddNeuroMed, two major Alzheimer's disease cohorts, on individual feature level and observed significant differences between them. These results raise concerns about the ability to generalize findings based on a single cohort dataset. Despite identified differences, we could validate our previously published, ADNI trained artificial intelligence model to predict Alzheimer's disease diagnosis (Khanna et al., 2018) on 244 AddNeuroMed subjects. Validation resulted in a high prediction performance of above 80% AUC (area under receiver operator characteristic curve) up to 6 years before Alzheimer's disease diagnosis. Propensity score matching identified a subset of patients from AddNeuroMed, which showed significantly smaller demographic differences to ADNI. For these patients an even higher prediction performance of 88% AUC was achieved. This implies that recruitment criteria and following cohort composition can indeed significantly affect model performance. In conclusion, our work highlights two aspects: It showcases a systematic approach for comparing clinical cohorts on feature level, and is one of the rare cases in the neurology field in which an artificial intelligence model was externally validated via independent cohort study data.

References: Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., Fröhlich, H. (2018). Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms. Scientific reports, 8(1), 11173.

5.2 Neural Ordinary Differential Equations for Modeling and Predicting Parkinson's Disease Progression

Marc Gómez-Freixa, Holger Fröhlich

Bonn-Aachen International Center for Information Technology

Neurodegenerative diseases such as Parkinson's Disease (PD) show a highly heterogeneous and complex progression behavior. Methods to predict the longitudinal and multivariate symptom spectrum of PD on an individual patient level could allow for identifying better targeted therapies in the future.

In this work we combined recurrent neural networks with latent ordinary differential equations (Neural ODEs - Chen et al., NIPS 2018). More specifically, we adapted and modified the original approach by Chen et al. by including implicit missing value imputation. Moreover, we combined demographic and genetic information with time dependent multivariate symptom measurements.

Extensive comparison of our method to conventional recurrent neural networks (LSTMs) as well as generalized linear mixed models show a competitive prediction performance. Overall our work demonstrates the possibility to combine mechanistic and purely data driven modeling strategies.

5.3 A bootstrap approach to estimate false positives in viral meta genomics

Babak Saremi, Klaus Jung

Institute for Animal Breeding and Genetics, University of Veterinary Medicine of Hannover

BACKGROUND:
The constant improvement of high-throughput DNA sequencing technologies lead to a rapidly increasing amount of genomic sequences that are available in public databases. These genome sequences are used in a variety of bioinformatic analysis. One of those bioinformatic applications aims to identify viral genomic sequences from an infected host [1], by sequencing a biological tissue sample from the host. However, such sequence data can be affected by different technical errors, e.g. probe preparation, false base calling etc., leading to wrongly identified viruses. Repeatedly sequencing of the sample from the infected host for the purpose of judging the robustness is too costly.

METHODS:
Here, we present a bootstrap approach for re-sampling sequencing data to approximate the robustness of NGS experiments in viral meta genomics. The bootstrap algorithm is implemented in python and draws Phi amounts of sequences from a fast-Q file. This sampling procedure is repeated B times and yields B bootstrap fast-Q files. Each fast-Q file is then mapped against a reference database of viral sequences and at the end, all B results are summarized. To evaluate our new approach, the bootstrap procedure is run on a simulated set of paired-end fast-Q files, using the Art-tool [2], with known viral sequence content. This simulated set drawn on the basis of one pig chromosome to represent a host and 100 selected viral genomes.

RESULTS & DISCUSSION:
Using the proposed bootstrap approach we could derive intervals for the number of reads mapped to each of the 100 virus genomes from which sequences were spiked-in into the fast-Q-file from the host. These intervals can be used as a measure to judge the robustness of the findings. Currently, we are further developing the approach to make it applicable for the reduction of false positives.

References:

1 Kruppa, J., Jo, W. K., van der Vries, E., Ludlow, M., Osterhaus, A., Baumgaertner, W., & Jung, K. (2018). Virus detection in high-throughput sequencing data without a reference genome of the host. Infection, Genetics and Evolution, 66, 180-187.

2 Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2011). Art: a next-generation sequencing read simulator. Bioinformatics, 28(4), 593–594.

5.4 Integrative analysis of peripheral N-acetylaspartate metabolism

Polina Lakrisenko

Helmholtz Zentrum München

Canavan disease (CD) is a rare leukodystrophy caused by mutations in the ASPA gene, leading to severe neurodegeneration and short life expectancy. To date, the exact disease mechanism is poorly understood and therapeutic options are scarce. The ASPA gene encodes aspartoacylase, an enzyme catalyzing the degradation of N-acetylaspartate (NAA). Despite NAA being the second-most abundant metabolite in the mammalian brain, its functional role is poorly understood. Recently, several independent groups found NAA metabolism to also play important roles in non-nervous tissues, e.g. in adipocytes, immune cells, lung and prostate cancer cells, pointing towards a previously overlooked wide relevance of peripheral NAA metabolism.

Together will clinical and experimental partners, we are generating a computational model around NAA metabolism and infer model parameters from stable isotope-assisted metabolomics data, RNA-seq data, and other experimental data. We will use this model to iteratively generate and test various hypotheses around the NAA metabolism and its function. Thereby, we expect to obtain a more profound understanding of the roles of NAA in metabolism and signaling, particularly in cells outside the brain, which is relevant for understanding and treatment of CD and other diseases.

5.5 Stratifying PD patients by disease progression using advanced machine learning techniques

Ashar Ahmad

University of Bonn, B-IT

Parkinson Disease (PD) is a neurodegenerative disorder known to affect over 10 million people worldwide. Patients suffering from PD exhibit a broad range of signs and symptoms ranging from cognitive decline to impaired motor activities. The progression pattern of the PD patients tends to be of a highly heterogeneous nature. Stratifying PD patients would potentially allow us to identify patient subgroups more likely to respond to drugs in clinical trials hence allowing for a reduced sample size in the trial. Moreover, knowledge of the important features which lead to faster/ slower progression from our ML approach might be useful in future drug discovery for certain PD subgroups. In the present work we use data from The Parkinson's Progression Markers Initiative (PPMI) which is a popular longitudinal clinical study to comprehensively evaluate PD patients in terms of multi-modal data (imaging, genetics, CSF biomarker, clinical and behavioral assessments). We use this data set to build a machine learning model (multivariate functional clustering) which identifies three distinct patient subgroups based on their multi-modal longitudinal progression patterns. Subsequently we use these three cluster labels to build a classification model based on only the baseline data, along with genomic and imaging profiles with the goal to predict these three progression sub-populations. For this purpose, we compare ensemble based supervised learning approach (Random Forest) with Penalized Generalized Linear Model approach (Group LASSO). We report a cross-validation performance of around 70% AUROC which increases to more than 80% AUROC when we include data after one-year follow up. As a final validation step we use the Parkinson's Disease Biomarkers Program (PDBP) data as external validation data set. We confirm the existence of the previously established three patient- subgroups in the PDBP data by predictions obtained from our multivariate functional classifier.

5.6 Simultaneous inference of gene association networks and cell types from single-cell RNA sequencing data

Stefan Schrod

Universität Regensburg

High-throughput gene-expression data can be used to study stochastic biological processes. These processes can be analysed using probabilistic graphical models (PGMs). PGMs assume an underlying probability distribution, where the parameters give the conditional (in-)dependency structure of gene-expression data. We propose a method that simultaneously identifies cell-populations and infers the corresponding PGMs. Our Method is based on the Graphical Lasso with penalty terms that enforce sparse network differences. It simultaneously identifies the underlying cellular populations using the expectation maximization (EM) algorithm.

First, we validate our method in simulation studies, where we compare different choices for the penalty terms, such as Lasso, fused and group Lasso terms. We finally apply our method to single-cell RNA sequencing data of T cells in hepatocellular carcinomas. Here, we compare the estimated cell populations with results from FACS sorting.

In summary, we present a method that simultaneously learns cell populations and infers the underlying gene-gene association networks. It allows to define new cell populations and simultaneously provides its main differences compared to established cell populations.

5.7 Statistical Modeling Approaches to Predict Functional and Cognitive Decline for Alzheimer's Disease Patients

Meemansa Sood

Fraunhofer SCAI

Alzheimer's Disease (AD) is associated with staggering costs which are particularly due to the impact of caregiving for increasing dementia population. Typically, the tests that are conducted on AD subjects rely on direct observations or caregiver recall which follows the principle of "diagnose and treat". However, in order to reduce such extensive costs, today's research is shifting towards the use of digital technologies such as wearables and home-based monitoring devices. These methods are trying to shift the focus from "diagnose and treat" to "predict and pre-empt". The studies focusing on these methods have a limitation that they have a limited observation time range and follow a cross-sectional framework. In order to overcome this limitation, we need to map the measures from these devices to the functional and cognitive decline of patients measured in a longitudinal framework.

Our research focuses on application and comparison of different statistical modeling approaches on functional and cognitive measures obtained from heterogenous longitudinal cohorts. We identified biosignatures that were indicative of changes in functional and cognitive status and applied multiple statistical modeling approaches. Foremost approach focused on learning spatiotemporal patterns[1] in order to produce individualized models and second approach focused on linear mixed models. These modeling methods helped us to prioritize functional and cognitive measures in patients in a temporal order. The pattern of trajectories along the progression of disease in the datasets used were observed to be similar. These longitudinal trajectories obtained will be further mapped to the digital readouts from cross-sectional study in order to predict functional and cognitive decline in patients at an early onset of the disease. This analysis will further assist in patient-specific prediction of the trajectories for the major functional outcomes captured by the wearable devices, such as quality of life, sleep quality, motor function, and activities of daily living.

1. https://hal.inria.fr/hal-0196482

5.8 Simulation of cancer derived extracellular vesicles metabolism

Miroslava Cuperlovic-Culf

National Research Council of Canada, Canada

Extracellular vesicles, exosomes, are cell-derived packages of proteins, nucleic acids, metabolites and lipids. They have an important role in cell-cell communication and modifications of cellular environment with specific roles in tumour metastasis. At the same time, as specific packages of biological material they are increasingly studied as diagnostic tools as well as carriers of drug and gene therapies. In all of these roles exosomes have to endure and adapt to different environments while circulating in body fluids. Their ability to adjust and function will depend on their lipid bilayer, their metabolic content as well as their enzyme and transport proteins. Knowledge of exosomes metabolic characteristics and adaptability is essential for their utilization as both diagnostic and treatment tools. In this work we have investigated possible flux of metabolic processes in exosomes derived from Glioblastomas (GBMs) using published proteomics information. Possible enzymatic reactions in exosomes have been explored against the NMR metabolomics data for GBM derived exosome. Machine learning analysis of exosome metabolomics as well as flux simulation model allowed us to determine major points in exosome metabolic process including both enzymatic reactions and transfer through exosome membrane and its dependence on external media. This result will be discussed in the context of GBM diagnostics and treatment delivery using exosomes.

5.9 Towards a mouse pneumonia atlas applying single cell RNA sequencing

Holger Kirsten

Uni Leipzig / IMISE

Single cell RNA sequencing (scRNAseq) is a core technology to obtain improved understanding of biological processes locally involving several cell-types. Here, we show results from our collaboration comparing scRNAseq experiments in mice of different genetic background with and without pneumonia due to different infectious bacterial strains. About 5000 cells per single experiment were analyzed applying 10X Genomics single cell gel beads in emulsion technology, and sequenced on an Illumina NextSeq. Expected cell types include e.g. alveolar epithelcells, endothelcells, smooth muscle cells, neutrophils, and alveolar macrophages. For analysis, we apply and compare different tools, such as seurat, scater, scanpy, cell ranger, and galaxy.

Understanding of gene expression patterns differing between those conditions as well as between relevant cell types will be a first step to a deeper systemic understanding with the ultimate goal of a mouse pneumonia atlas.

5.10 Deep learning for clustering of multivariate longitudinal clinical patient data with missing values

Johann de Jong

UCB Biosciences GmbH

In the literature, the problem of clustering multivariate short time series is still largely unaddressed. However, multivariate short time series are very common in the analysis of clinical data, when multivariate patient measurements are taken over time. The clustering (stratification) of such clinical data is additionally complicated by a typically high degree of missingness.

For this purpose, we developed a deep learning–based method, variational deep embedding with recurrence (VaDER). VaDER relies on a Gaussian mixture variational autoencoder framework, which is further extended to (i) model multivariate time series using long short term memory cells (LSTMs) and (ii) directly deal with missing values by implicit imputation combined with loss reweighting.

We technically validated VaDER by accurately recovering clusters from simulated and benchmark data with known ground truth clustering and varying degrees of noise and missingness. We then used VaDER to successfully stratify (1) Alzheimer's disease patients and (2) Parkinson's disease patients into subgroups characterized by clinically divergent disease progression profiles, as measured by scores on a diverse range of cognitive and motor symptom assessments. Analyses of additional data (brain imaging, biomarkers, ...) furthermore demonstrated that these clinical differences reflected known underlying aspects of Alzheimer's disease and Parkinson's disease.

We believe our results show that VaDER can be of great value for future efforts in patient stratification, and multivariate short time series clustering in general.

5.11 A cell cycle dependent population dynamics model with parameter inference from scRNA-seq data

Simon Merkt

University of Bonn

Background:
In the last few years, single cell became widely used and hence the interest in analyzing these data mathematically also grew. Recently pseudodynamics - a method for investigating the temporal evolution- has been developed. It models a time-series of single cell snapshot data as if they were part of a single continuous trajectory.

Methods:
We reduce the highly dimensional data to a one dimensional space, the so called state space. The dynamics of individual cells can then be described by the changes in their cell state. The population dynamics are governed by a diffusion-advection-reaction equation. Its solution can be obtained numerically with the finite volume method and the necessary parameters for drift, diffusion and growth can be inferred from the data.

Results/Goals:
This model of the dynamics has been enhanced to consider also different cell cycle stages. Mathematically speaking, this leads to a circle system of partial differential equations where each equation is connected to the next one by a transition rate. The resulting partial differential equation system has been implemented numerically. Since the current stage of a cell can be obtained from scRNA-seq data, each cell can now be assigned to one of the equations and used for inference of its progression rate.

5.12 The more the better: How to gain knowledge from expression data enriched pathways

Tamara Raschka

Fraunhofer SCAI

The improvements in measuring gene expression data in less time and with decreasing costs have resulted in a tremendous increase of transcriptomic data during the last years. However, most of the data is produced for a single task, and will never be touched again. A huge amount of produced transcriptomic data, from microarray, as well as from next-generation sequencing technologies, is stored in the NCBI Gene Expression Omnibus (GEO). If a re-analysis of this data is done, mostly only a very small subset of GEO is used. But can't we do something big with these data?

What is also growing during the last years, are pathway databases. Because pathways are artificial concepts depending on a context in which they are developed, many different pathway databases are appearing. Also, the relations within pathways are not annotated with context and also do not have weights based on which pathways can be ranked for a specific sample or a dataset. Can't we improve information on pathways by linking additional data?

Bringing together the two worlds of gene expression data and pathways can be a key to gain more information from the data and to generate additional knowledge for pathways. Furthermore, the linking of both gives the possibility to add context to the relations within a pathway, weight the relations and rank pathways based on its variability across datasets. This can be achieved by, first, generating co-expression networks for GEO datasets and identification of invariant gene pairs. Clearly, building the co-expression networks with WGCNA will be a computationally intensive task, and therefore the workflow will be first tested for specific pathways, so the computation is limited down to a subset of genes only. The invariant edges of the network are common to all data, so they can be entitled as 'housekeeping'. Based on the variability of a pathway, so the variability of relations within the pathway, they can be ranked and assigned to a specific context. Whether the context is significant for a specific pathway will be tested by comparing pathway scores of 'all' data and the 'context' data against each other. In the end, each pathway, and every single relation of it, can be enriched with the context in which it is stable. This innovative approach, a 'Big Data' application, will allow identifying invariantly co-expressed gene pairs and labeling triples, within pathways, based on context diversity and variability. It will enhance the transparency and computability of pathways.