Program for Dagstuhl Seminar „Addressing the Computational Challenges of Personalized Medicine“

Niko Beerenwinkel, Holger Fröhlich, Franziska Michor, Susan Murphy

19.11., Sunday:

- 18:00: dinner

- 19:00 – 19:15: Welcome and Introduction (Holger Fröhlich)

- 19:15 – 21:00: Self intro (3 slides each):
    - What are my main research topics?
    - Which aspect of the seminar topic is particularly important to me and should be discussed within the seminar?
    - What are my expectations for the seminar?

- 21:00: Social get together

20.11., Monday:

Session 1: Enhancing Prediction Performance

- 9:00 – 9:40: Keynote talk (Yves Moreau): Bayesian matrix factorization with side information (incl. 15 minutes discussion)
- 9:40 – 10:20: Keynote talk (Kumar Santosh): Dynamic patient re-stratification Using Mobile Sensors (incl. 15 minutes discussion)
- 10:20 – 10:30: splitting into working groups (4 – 5 people)

- 10:30 – 11:00: coffee break

- 11:00 – 12:30: Working groups

- 12:30 – 14:00: Lunch break

- 14:00 – 15:00: Presentation and collection of results from working groups (collector: Niko Beerenwinkel)

- 15:00 – 15:30: coffee break

Session 2: Improving Interpretability

- 15:30 – 16:10: Keynote talk (Andreas Schuppert): Hybrid models - combining mechanistic and statistical modeling approaches (incl. 15 minutes discussion)
- 16:10 – 16:30: Keynote talk (Rudi Balling): Visualizing and integrating biological knowledge (incl. 15 minutes discussion)
- 16:30 – 16:40: splitting into working groups (4 – 5 people)

- 18:00: dinner, social get together

21.11., Tuesday

- 9:00 – 10:30: Working groups
- 10:30 – 11:00: coffee break
- 11:00 – 12:00: Presentation and collection of results from working groups (collector: Holger Fröhlich)

- 12:00 – 13:30: Lunch break

Session 3: Translation into Clinical Practice

- 13:30 – 14:10: Keynote talk (Michael Rebhan): Enhanced translation of multi-modal stratification models, as a basis for Precision Medicine (incl. 15 minutes discussion)
- 14:10 – 14:20: splitting into working groups (4 – 5 people)

- 14:20 – 14:50: coffee break

- 14:50 – 16:20: Working groups

- 16:20 - 16:50: coffee break
- 16:50 – 17:50: Presentation and collection of results from working groups (collector: Susan Murphy)

- 18:00 dinner, social get together

22.11., Wednesday

- 9:00 – 10:00: Wrap up, report drafting, discussion of perspectives paper
- 10:00 – 11:00: Planning of next steps (publications, grants, scientific ideas, community + political efforts, …)

Talk Abstracts

Bayesian matrix factorization with side information

Yves Moreau - University of Leuven, Belgium

Matrix factorization/completion methods provide an attractive framework to handle sparsely observed data, also called “scarce” data. A typical setting for scarce data are is clinical diagnosis in a real-world setting. Not all possible symptoms (phenotype/biomarker/etc.) will have been checked for every patient. Deciding which symptom to check based on the already available information is at the heart of the diagnostic process. If genetic information about the patient is also available, it can serve as side information (covariates) to predict symptoms (phenotypes) for this patient. While a classification/regression setting is appropriate for this problem, it will typically ignore the dependencies between different tasks (i.e., symptoms). We have recently focused on a problem sharing many similarities with the diagnostic task: the prediction of biological activity of chemical compounds against drug targets, where only 0.1\% to 1\% of all compound-target pairs are measured. Matrix factorization searches for latent representations of compounds and targets that allow an optimal reconstruction of the observed measurements. These methods can be further combined with linear regression models to create multitask prediction models. In our case, fingerprints of chemical compounds are used as “side information” to predict target activity. By contrast with classical Quantitative Structure-Activity Relationship (QSAR) models, matrix factorization with side information naturally accommodates the multitask character of compound-target activity prediction. This methodology can be further extended to a fully Bayesian setting to handle uncertainty optimally, which is of great value in this pharmaceutical setting where experiments are costly. We have developed a significant innovation in this setting, which consists in the reformulation of the Gibbs sampler for the Markov Chain Monte Carlo Bayesian inference of the multilinear model of matrix factorization with side information. This reformulation shows that executing the Gibbs sampler only requires performing a sequence of linear regressions with a specific noise injection scheme. This reformulation thus allows scaling up this MCMC scheme to millions of compounds, thousands of targets, and tens of millions of measurements, as demonstrated on a large industrial data set from a pharmaceutical company. We have implemented our method as an open source Python/C++ library, called Macau, which can be applied to many modeling tasks, well beyond our original pharmaceutical setting , see https://github.com/jaak-s/macau/tree/master/python/macau.

Dynamic Patient Restratification Using Mobile Sensors

Kumar Santosh - University of Memphis, USA

Recent advances in wearable sensing and mobile computing have opened up unprecedented opportunities to quantify dynamic changes in an individual’s health state as well as key physical, biological, behavioral, social, and environmental factors that contribute

to health and disease risk, anytime and anywhere. For example, smart watches can not only track physical activity, but they can also be used to monitor stress (from pulse rate), eating, brushing, driving, and smoking behaviors (from hand gestures). By simultaneous monitoring of changes in health status, exposures to surrounding geographical, environmental, visual, social, and digital worlds, and personal behaviors (both risky and healthy), mobile health (mHealth) can help discover new predictors of health outcomes.
By monitoring the exposure to these health risk predictors, mobile health offers an opportunity to introduce temporal precision in precision medicine, especially when mHealth data is used together with traditional sources of biomedical data (e.g., genomics, clinical). Longitudinal nature of mHealth data and the fact that it comes from the natural free-living environment allows dynamic decision making such as adapting the treatments and interventions so as to maximize the efficacy and optimize the timing of delivery. Continuous monitoring of the context surrounding the individual and monitoring of the compliance and response to treatments and interventions offers additional opportunities for dynamic optimizations in a human-in-the-loop model.
Realizing these potential presents a rich multi-disciplinary research agenda. It includes sensor design and mobile system design for optimizing data collection with minimum user burden, mobile sensor big data modeling to convert voluminous mobile sensor data into actionable information, sensor-triggered intervention modeling that leverages dynamic optimization opportunities to discover the most efficacious and temporally-precise treatments and interventions, and engaging visualizations to encourage health and wellness-supporting daily behaviors using new insights gained from mHealth data.

Hybrid models - combining mechanistic and statistical modeling approaches

Andreas Schuppert - RWTH Aachen, Germany

Modeling for personalized medicine requires methods enabling to predict reliably the evolution of the diseases, the response on therapies as well as the therapeutic adverse side effects for individual patients. However, due to a lack of understanding of the broad range of mechanisms affecting diseases and therapies, pure mechanistic modeling rarely results in satisfactory precision.
On the other side, pure machine learning – based modeling methods are hampered by their conceptually high data demand for model training and their lack of extrapolation. In patient populations, the intrinsic mutual control loops inside the system “patient” in combination with the high variety of optional covariates result in statistically poor, biased distributions of data in high dimensional data spaces, hampering machine learning even in large “real world evidence” data sets.
Hence, a combination of mechanistic and machine learning in a hybrid model is required in order to achieve the necessary precision of the models. Hybrid modeling had been developed for chemical and biotechnological engineering in order to tackle the lack of process data, combined with common lack of quantitative understanding of the reaction kinetics . The mathematical basis of data representation by means of hybrid models goes back to Hilbert’s famous 13th problem and has been intensively discussed by Kolmogoroff, Arnold and Vitushkin . Later it could be shown that the knowledge of the true system

structure without any mechanistic knowledge is sufficient to break the curse of dimensionality, to reduce the data demand for model training and to enable extrapolability of the models . The inverse problem, namely the identification of model structures from data, has recently been discussed in the context of systems biology .
These results apparently have a strong relationship to the current development of deep learning technologies. We expect that a future integrative technology might result in a modeling platform satisfying the requirements of personalized medicine.

Visualizing and Integrating Biological Knowledge

Rudi Balling - University of Luxemburg, Luxemburg

The LCSB is engaged in a number of community efforts to develop novel tools for the visualization, annotation and integration of network-encoded knowledge in biomedical research. In order to capture the rapidly increasing information and inter-relationships between different factors contributing to Parkinson´s disease (PD), we have established a “PD-map”. This map is a manually curated knowledge repository and serves as a computationally tractable representation of all known molecular interactions involved in the pathogenesis of Parkinson´s disease. The disease map offers research-facilitating functionalities such as the overlay of experimental data and the identification of drug targets on the map. A major effort is also geared towards the development of genome-scale human and human gut metabolic reconstructions integrating the full spectrum of metabolic and transport reactions that can occur in a given organism. The goal is to develop a comprehensive knowledge base of human metabolism integrating pharmacogenetic associations, large-scale phenotypic data and structural information for proteins and metabolites.

Enhanced translation of multi-modal stratification models, as a basis for Precision Medicine

Michael Rebhan - Novartis, Switzerland

Progress in Precision Medicine and Personalized Health is linked to our ability to translate increasingly complex ‘multi-modal stratification’ models from discovery to validation, and finally to real world healthcare settings where they can generate impact on patient outcomes. Such models need to be able to computationally deal with a diversity of signals from an increasing number of ‘channels’ that can influence stratification, including those derived from molecular biomarkers, imaging technology, and ‘digital biomarkers’, to name a few. Such models would, down the road, help us predict not only the best intervention for a particular patient, but also the best time and context for delivering it, considering disease progression knowledge, patient needs and priorities, and different healthcare settings. In this session, we will discuss the idea of co-designing an open innovation ecosystem for community-based learning on such models, ‘on top’ of the current health data silos. As there are many challenges on the translational path for these models, we will discuss potential solutions to

explore as a community. How to best conduct high quality clinical validation studies that can help to bridge the gap between early research and responsible first use of multi-modal stratification models in clinical decision making? How could outcome-based feedback loops help with community-based learning, beyond the clinical institutions involved in patient care? How can open learning ‘on top of the data silos’ look like, in practice? As we discuss those challenges, we will try to consider the full complexity of the health innovation landscape with its many stakeholders (patients, physicians, payers, basic / applied researchers, regulators etc.), and real life challenges, as this will help us co-design meaningful translational paths. In addition, we will discuss guiding principles that can help with the community build, e.g. transparency (of data and algorithms), and their role in such an effort.