Report from Dagstuhl Seminar 17472

Addressing the Computational Challenges of Personalized Medicine

Edited by

Niko Beerenwinkel[1], Holger Fröhlich[2], and Susan A. Murphy[3]

1 **ETH Zürich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland**
niko.beerenwinkel@bsse.ethz.ch

2 **UCB Biosciences GmbH, Alfred-Nobel-Str. 10, 40789 Monheim, Germany and University of Bonn, Bonn-Aachen International Center for IT (b-it), Dahlmannstr. 2, 53127 Bonn, Germany**
holger.froehlich@ucb.com; frohlich@bit.uni-bonn.de

3 **Harvard University, Department of Statistics, Science Center 400 Suite, One Oxford Street, Cambridge, MA 02138-2901, USA**
samurphy@fas.harvard.edu

Abstract

This report provides an overview of the talks and the working group reports from the Dagstuhl Seminar 17472 "Addressing the Computational Challenges of Personalized Medicine". The seminar brought together leading computational scientists with different backgrounds and perspectives in order to allow for a cross-fertilizing and stimulating discussion. It thus joined expertise that is usually scattered in different research communities. In addition, selected medical researchers, pharmacogenomics researchers and behavioral scientists provided their input and established the link of the computational to the more medical aspects of personalized medicine (PM). The talks and corresponding discussion spanned mainly three areas: 1) how to enhance prediction performance of computational models for PM; 2) how to improve their interpretability; 3) how to validate and implement them in practice.

Seminar November 19–22, 2017 – http://www.dagstuhl.de/17472

1998 ACM Subject Classification I.2.6 Learning, I.6 Simulation and Modeling, J.3 Life and Medical Sciences

Keywords and phrases data science, machine learning, computational modeling, bioinformatics, systems biology

Digital Object Identifier 10.4230/DagRep.7.11.130

1 Executive Summary

Niko Beerenwinkel
Holger Fröhlich
Susan Murphy

Personalized medicine (PM) is understood as a non-traditional medical approach, in which patients are stratified based on their disease subtype, disease risk, disease prognosis or treatment response using specialized diagnostic tests. High promises for the whole health care sector are associated with PM, and correspondingly the topic has received a lot of

attention during the last years. PM is tightly connected to and dependent on computational sciences (computer science, mathematical modeling, computational statistics, bioinformatics). Currently, shortcomings of computational methodology constitute an important bottleneck for PM, which hinders full realization.

The goal of the planned seminar was to bring together an international and interdisciplinary group of experts in different computational science disciplines in order to discuss, how some of the major existing computational challenges could be better addressed in the future, namely:

1. how to enhance prediction performance of computational models for PM
2. how to improve their interpretability
3. how to validate and implement them in practice

The seminar joined together expertise that is usually scattered across different disciplines. The seminar had a strict focus on computational methodology, but few selected non-computational scientists closed the gap to the application field.

2 Table of Contents

3 Overview of Talks

3.1 Bayesian matrix factorization with side information

Yves Moreau (University of Leuven – Leuven, Belgium, moreau@esat.kuleuven.be)

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. https://github.com/jaak-s/macau/tree/master/python/macau.

3.2 Dynamic Patient Restratification Using Mobile Sensors

Kumar Santosh (University of Memphis – Memphis, USA, santosh.kumar@memphis.edu)

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.

3.3 Causality

Marloes Maathuis (ETH Zürich – Zürich, Switzerland, maathuis@stat.math.ethz.ch)

Causal questions are fundamental in all parts of science. Answering such questions from non-experimental data is notoriously difficult, but there has been a lot of recent interest and progress in this field. I have discussed current approaches to this problem and have outlined their potential as well as their limitations.

3.4 Hybrid models – combining mechanistic and statistical modeling approaches

Andreas Schuppert (RWTH Aachen – Aachen, Germany, schuppert@aices.rwth-aachen.de)

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.

3.5 Visualizing and Integrating Biological Knowledge

Rudi Balling (University of Luxemburg – Luxemburg, rudi.balling@uni.lu)

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.

3.6 Computational Analysis of Viral Drug Resistance

Thomas Lengauer (Max-Planck Institute for Informatics – Saarbrücken, Germany, lengauer@mpi-inf.mpg.de)

We present a concrete case of translational research in computational biology in 4 steps. The problem is to estimate the resistance to HIV to individual drug based on viral genotype and to combinatorial therapies with respect to their estimated effectiveness.

1. gene2pheno[resistance] estimates the level of resistance of HIV to individual inhibitors of viral protease and reverse transcriptase. The software has first been online 15 years ago. It is used in clinical practice, but has strong competition from expert rule systems. The software interprets its predictions in terms of effect of individual mutations in the HIV genome.
2. gene2pheno[coreceptor] Our "blockbuster" estimates viral coreceptor usage It provides a significance estimate. There is no competition from rule based systems.
3. gene2pheno[THEO] ranks combinations of drug therapies w.r.t. estimated effectiveness. In contrast to the previous tools this one has not entered clinical routine, partially because the predictions are not interpreted.
4. g2p2 is our new development that is aimed at bringing therapy predictions to clinical routine. It merges mathematical analysis with traditional schemes of therapy composition and is interactive.

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

Michael Rebhan (Novartis AG – Basel, Switzerland, michael.rebhan@novartis.com)

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.

4 Discussed Challenges and Possible Approaches

Author: Niko Beerenwinkel, Holger Fröhlich, and Susan Murphy

The following comprises a summary of the problems and possible approaches that were discussed in different working groups and within the panel.

Enhancing Prediction Performance

Performance Metrics

It is necessary to consider performance metrics apart from the established area under ROC curve (AUC). The choice of performance metric should depend on the actual prediction problem at hand.

Data Quality and Systematic Biases

Data quality is one of the reasons behind low prediction performance. Data quality is a continuous concern, specifically with respect to omics data. Robust loss functions in machine learning methods should be considered.

Prediction performance is also affected, if the data represents unknown mixtures of different biological origin. For example, tumor biopsies often contain a mixture of actual tumor and stroma cells, which impacts measured gene expression. A possible approach is to de-convolute the original data via mixture (regression) models. At this point a Gaussian assumption for transcriptomics data seems feasible.

It is known that independently collected patient cohorts exhibit a systematic difference in their expression profiles to the original training cohort. The recently introduced zero-sum regression approach is a way to address this issue [1].

Feature Engineering and Extraction

Feature engineering remains a crucial topic for successful modeling, because it allows for using prior knowledge. Such prior knowledge could also come from similar data, which has been collected for different purposes.

In addition to feature engineering, extraction of (latent) meta-features is likely to be a successful strategy. Methods include matrix factorization techniques as well as auto-encoder networks.

Use of Multi-Modal Data

Multi-modal, longitudinal data is widely believed to provide a more detailed view on the complex relationship between biology and clinical outcome, which we try to capture with models in personalized medicine. Multi-modal patient trajectories are possibly embedded into a lower dimensional latent space, in which patterns become more obvious than in the original space. Matrix factorization approaches might be one way to identify a suitable latent space.

Despite of a multitude of available methods multi-modal data integration remains a challenge, specifically when fusion of static (e.g. genomic) and longitudinal data (e.g. clinical features) is desired. In the data science literature early, intermediate and late integration

schemes are discussed, which have all their advantages and disadvantages. The optimal data fusion strategy is always data dependent and has thus to be found empirically.

In the future, data retrieved from web mining (patient blogs, social media) could play an increasing role. It might be possible to use these data within a Bayesian learning scheme, e.g. to define informative priors.

Improving Interpretability

Disease Maps

Disease maps describe cause-effect relationships between multi-modal molecular and clinical data entities. Disease maps are not computational models per se, but could be used in two different ways to obtain better interpretable prediction models:

- Post-hoc analysis of features in the model, e.g. via enrichment analysis and variants thereof.
- Embedding of network information into feature selection.

Both approaches are established in principle, but may require further adaptation to a specific problem, e.g. by defining subsets of the disease map, or extracting and engineering of appropriate features based on data.

Disease maps could also help informing causal network inference (see next paragraph).

Causal Models

Predictive models are often highly complex and not necessarily causal, which hinders the acceptance by physicians and limits scientific insights into the underlying pathophysiological mechanisms. Judea Pearl has established a widely accepted theory of causality in the context of probabilistic graphical models [2]. However, the graph structure of causal models can in general only be identified from observational data up to equivalence classes. Nonetheless, it is possible to predict bounds of intervention effects from purely observational data under certain conditions [3]. There is a need to better evaluate these methods in the context of personalized medicine. It has to be checked, how reproducible the results are and whether causal network models could inform data collection (when and what to collect) in the future. Moreover, the exact conditions under which for an individual prediction of causal intervention effects are possible should be clarified.

Hybrid Models

Causal models are not necessarily pointing towards detailed biological mechanisms. On the other hand fully mechanistic models are limited by the available background knowledge, which is often incomplete. Hybrid models combine partially available quantitative mechanistic and machine learning models into one unified framework. Within that framework machine learning "black-boxes" detect and correct errors in the mechanistic part. Black-box and mechanistic models can be integrated into a hybrid network and trained via mathematical optimization methods. Hybrid models currently have a theory gap. Nonetheless, the approach has been used successfully by Andreas Schuppert and colleagues for predicting drug response in diabetes I (unpublished work): This was possible, because there is a mechanistic model for diabetes I available. Behavioral aspects (eating, exercise, etc), which are also important for the disease, can be put into the black-box model part. There is much less known about diabetes II and consequently, there is no mechanistic model.

Critical Transitions

The development of many diseases may be interpreted from the perspective of a phase transition in a dynamical system. In physics this phenomenon is well known and appears in many models. Reliable detection of phase transitions in disease development could ultimately help to detect mechanistic biomarkers in the future and enable early disease diagnosis and prevention. There are hints in the literature that phase transitions might be detectable from data via simple statistical methods, such as variability and correlation. However, more evaluation is needed to demonstrate the actual utility in the context of personalized medicine.

Enhancing Transition to Clinical Practice

Better Transparency and Interpretability of Models

The lack of interpretability and transparency is one of the key obstacles that hinders acceptance of machine learning models by physicians and regulatory agencies. The pure focus on prediction performance is misleading. Additional measures such as stability, enrichment of existing knowledge and cross-study applicability should be considered. Moreover, there is the need to link model predictions with a "narrative" that can be understood by doctors and patients. Such a narrative may be generated in different ways:

- by visualizing molecular features that drive the prediction, possibly also with the help of disease maps
- by showing different decision alternatives together with their confidences whenever possible
- by showing and visualizing close patients from the training data (w.r.t. some metric)
- by generating a medical report for each patient in an automated fashion
- by linking latent model features back to biological knowledge

The latter point will require further research, but one possibility might be to look for evolutionary conserved disease modules. This could at the same time open the door to better utilize animal models.

Continuous Updating of Prediction Algorithms

There is a need for a clear and transparent process for continual iteration/updating and revalidation for precision medicine software tools. The notion of CLIA (Clinical Laboratory Improvement Amendments) labs provide a template for how health-related software tools (diagnosis, prediction, decision support) developed to inform precision medicine can be validated and re-validated in an clear, transparent manner as the tool is continually updated. CLIA labs are certified labs that go through a process of regular re-certification and monitoring by FDA and other regulatory agencies in the US. These labs follow a SOP, e.g. an approved, transparent and documentation process. Currently CLIA labs monitor medical devices, which can include software diagnosis tools. When a CLIA lab waives certification the tool can be used in practice. Most importantly the developer of the tool can update the software tool. However the CLIA labs are independent and decide when they will re-validate a software tool (so maybe not each time you upload a new version).

5 Conclusions

The current hype around AI and machine learning has to be contrasted with the reality, in which we are facing a number of challenges in the context of personalized medicine. These range from insufficient prediction performance over lack of model interpretability up to difficulties to engage people moving further on to clinical practice with a given model.

The current machine learning hype is dangerous, because it rises inappropriate expectations. In order to manage these expectations there is the strong need to better inform physicians about the current opportunities, challenges and future potential of data science in medicine. We therefore plan to publish a paper in a medical journal focusing on the key learnings from this Dagstuhl seminar.

The overall vision formulated in the seminar was to enable a causal treatment of patients with the right drug at the right time. We see a number of intermediate steps towards this grand vision:

- high dimensional causal graphical models
- hybrid models
- understanding of critical transitions
- better use of the principles of evolution, e.g. by looking for evolutionary conserved disease modules

All of these steps are computational. This underlines the crucial relevance of computational models for enabling personalized medicine.

References

1 M. Altenbuchinger, T. Rehberg, H. U. Zacharias, F. Stämmler, K. Dettmer, D. Weber, A. Hiergeist, A. Gessner, E. Holler, P. J. Oefner, R. Spang, Reference point insensitive molecular data analysis, Bioinformatics, 33(2): 219 – 226, 2017

2 J. Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, 2000

3 M. Maathuis, D. Colombo, M. Kalisch, P. Bühlmann, Predicting causal effects in large-scale systems from observational data, Nat. Methods, 7(4): 247 – 48, 2009

Participants

- Rudi Balling
University of Luxembourg, LU
- Niko Beerenwinkel
ETH Zürich – Basel, CH
- Holger Fröhlich
UCB Biosciences GmbH – Monheim, DE
- Oliver Kohlbacher
Universität Tübingen, DE
- Santosh Kumar
University of Memphis, US
- Thomas Lengauer
MPI für Informatik – Saarbrücken, DE
- Marloes Maathuis
ETH Zürich, CH
- Yves Moreau
KU Leuven, BE
- Susan A. Murphy
Harvard University – Cambridge, US
- Teresa Przytycka
National Center for Biotechnology – Bethesda, US
- Michael Rebhan
Novartis AG – Basel, CH
- Hannes Röst
University of Toronto, CA
- Andreas Schuppert
RWTH Aachen, DE
- Matthias Schwab
IKP – Stuttgart, DE
- Rainer Spang
Universität Regensburg, DE
- Daniel Stekhoven
ETH Zürich, CH
- Jimeng Sun
Georgia Institute of Technology – Atlanta, US
- Andreas Weber
Universität Bonn, DE
- Daniel Ziemek
Pfizer GmbH – Berlin, DE
- Blaz Zupan
University of Ljubljana, SI