DAGSTUHL
REPORTS

Volume 5, Issue 8, August 2015

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at http://www.dagstuhl.de/dagpub/2192-5283

Publication date
January, 2016

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `http://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.5.8.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
http://www.dagstuhl.de/dagrep

Report from Dagstuhl Perspectives Workshop 15342

Power-Bounded HPC Performance Optimization

Edited by

Dieter Kranzlmüller[1] and Barry L. Rountree[2]

1 **Ludwig-Maximilians-Universität (LMU) & Leibniz Supercomputing Centre (LRZ), München, Germany,** kranzlmueller@lrz.de

2 **Lawrence Livermore National Laboratory (LLNL), Livermore, USA,** rountree@llnl.gov

Abstract

This report documents the program and the outcomes of Dagstuhl Perspectives Workshop 15342 "Power-Bounded HPC Performance Optimization". The workshop consists of two parts. In part one, our international panel of experts in facilities, schedulers, runtime systems, operating systems, processor architectures and applications provided thought-provoking and details insights into open problems in each of their fields with respect to the workshop topic. These problems must be resolved in order to achieve a useful power-constrainted exascale system, which operates at the highest performance within a given power bound. In part two, the participants split up in three groups, trying to address certain specific subtopics as identified during the expert plenaries. These subtopics have been discussed in more detail, followed by plenary sessions to compare and synthesize the findings into an overall picture. As a result, the workshop identified three major problems, which need to be solved on the way to power-bounded HPC performance optimization.

Perspectives Workshop August 16–21, 2015 – http://www.dagstuhl.de/15342

1998 ACM Subject Classification C.5.1 Super (Very Large) Computers, C.4 Performance of Systems, C.1.4 Parallel Architectures

Keywords and phrases Exascale computing, Energy efficiency, Power awareness, Scalability

Digital Object Identifier 10.4230/DagRep.5.8.1

1 Executive Summary

Barry L. Rountree
Dieter Kranzlmüller

The Dagstuhl Perspectives Workshop 15342 "Power-Bounded HPC Performance Optimization" has been an interesting experience, as in contrast to other workshops, we focused on the unknown characteristics of future exascale systems rather than on the state-of-the-art of todays petascale architectures. In order to do this, a large fraction of the workshop was spent on in-depth discussions in three working groups, while plenary sessions served to provide impulses on specific topics and to synthesize the findings of the breakout sessions. The key ingredient of this workshop has been the interaction between the particpants, leading to several new collaborations across vendors, national laboratories and academia.

The key findings of the workshop can be identified as follows:

- Power-bound performance optimization has different objectives according to the respective targets and operational goals. While infrastructure providers are often bound to a specific

spending, users want to utilize a resource at the maximum of its capabilities. As a result, any power-bound optimization must address multiple criteria, and the solution is rarely straight-forward but specific for a given setting.
- The currently available information on each layer of the computing environment is insufficient. Both, the availability of information with respect to its power characteristics, as well as the exchange between different layers, needs to be improved in order to optimize the operation of infrastructures and the execution of applications on a given system.
- Due to the number of dependencies, any optimization needs to find a good balance between "user happiness", total costs, and performance. These characteristics are imporant for both, providers and users, and a careful balancing strategy needs to be implemented without harming any interests of the actors too much.

The discussions at the Dagstuhl Perspectives Workshop have lead to the identification of a number of technical problems, which need to be addressed in the near future before achieving optimal results in a power-bound environment. As a conclusion, the participants agreed that a strategic and tactical agenda is needed, which identifies the individual problems and technologies as well as their interconnections, such that future systems can utilize this knowledge for new approaches of power-bound HPC performance optimization. The results of this investigations should be made available as a white book, which describes the strategy for future exascale systems.

2 Table of Contents

3 Overview of Talks

3.1 Introductory Remarks & Motivation

Barry L. Rountree (LLNL – Livermore, US)

The US Department of Energy and other supercomputing stakeholders believe that future high performance machines will be power limited, with a bound of 20 Megawatts suggested for the first exascale system. In this workshop we will explore the implications for power-limited computing, focusing primarily on optimization strategies. In particular, we will be making the distinction between energy-limited and power-limited systems, and discussing how hardware overprovisioning can increase performance for both.

3.2 Musings on Power, Programming Models, and Applications

David Richards (LLNL – Livermore, US)

In this talk we discuss how programming models for HPC applications might contribute to power optimization. Task-bases and other asynchronous programming models offer some hope in this regard as they expose concurrency as data dependencies in ways that might allow a runtime system to reorder or otherwise manage work to satisfy a power constraint. Unfortunately, no such models are production ready and it is unclear whether these models will be able to match the performance of more traditional HPC programming models. We discuss what developers might be willing to do so support power optimization. Finally, we hint at some of the challenges a runtime system might face in optimizing power by examining three examples of complex load imbalances that occur in real problems.

3.3 Open Problems in Processor Architecture

Jonathan Eastep (Intel – Hillsboro, US)

In this talk, I will discuss approaches for improving processor efficiency and tailoring processor architectures to work better with runtimes for optimizing system performance within a power bound. Approaches will include hardware-acceleration of basic building blocks of HPC codes, the addition of a 16-bit floating point format in SIMD units for the purpose of trading computational accuracy for additional performance, and optimization of the processor pipeline depth, transistor power-performance characteristics, and static power consumption to increase the utility of hardware over-provisioning strategies.

3.4 Performance Optimization vs. Power – Experiences with Petascale Earthquake Simulations on SuperMUC

Michael Bader (TU München, DE)

SeisSol is a high-order discontinuous-Galerkin software to simulate dynamic rupture and wave propagation processes during earthquakes. Working on unstructured meshes and applying static load distribution, it is representative for a large range of current simulation software. The high arithmetic intensity of its high-order discretization in space and time also make it an attractive candidate for peta- and maybe even exascale simulations. The talk specifically focused on power questions and issues that might arise with SeisSol simulations in the future:

- Measurements of the power consumption of SeisSol on the latest range of Intel CPU architectures revealed that optimising for time-to-solution implies improved energy-to-solution, already. Open questions include how to balance energy and time to solution in the choice of clock frequency and other hardware parameters, and how programmers could and should support runtime systems in this aspect.
- During the first petascale runs on the SuperMUC machine, SeisSol experienced machine crashes due to problems with the global power infrastructure, which were tracked down to strong variations and peaks in power consumption. As respective problems are likely to increase for exascale machines, will there be consequences for the software stack or even application programmers?
- Current processors already feature variations in their power consumption due to tolerances in the manufacturing process. Will such changes directly turn into performance variations in a power-bounded setup?

Will this make static load distribution, as currently applied in SeisSol, unfeasible on future machines? To conclude, the characterisation of the performance of simulation software will need to consider various quality numbers, especially time and energy to solution, and will open the question on how simulation software may interact with operating and runtime systems to mitigate power issues.

3.5 HPC Data Center Infrastructure Challenges Under A Power Bound

Torsten Wilde (LRZ – München, DE)

The era of energy efficient high performance computing (HPC) does not only create challenges for application developers and system software developers but also for the cooling infrastructure of HPC data centers. The move from air cooling to a mix of cooling technologies (air, indirect/direct water cooling, chiller supported and chiller-less cooling) in the data center coupled with the increasing dynamic power behavior of HPC systems makes the energy efficient operation of a data center nontrivial. This talk highlights current control challenges in the data center cooling infrastructure, using the LRZ data center as an example, and discusses how a power bound might help to improve the data center energy efficiency. We make the case that an adjustable (flexible) power bound might be beneficial in light of: the possibility of integrating renewable energy (mainly solar and wind power); changing

electricity costs when buying energy at the energy market; and changing outside conditions that effect the coefficient of performance (COP) of the data center cooling infrastructure. We discuss how a power bound can affect the four pillars (data center infrastructure, HPC system hardware, HPC system software, HPC applications) of the "4 Pillar model for energy efficient HPC data centers" and show that some connecting between all four pillars might be required. Finally, a concrete example of the possible benefit of a power bound is shown using data of the LRZ data center.

3.6 Future Directions in System Software on Power-Bounded Supercomputers

David K. Lowenthal (University of Arizona – Tucson, US)

System software play a significant role in power-bounded supercomputers. This talk covers possible future directions in system software.

4 Plenary and Breakout Sessions

The workshop was divided into a plenary track with the above mentioned keynote speeches, each covering specific aspects of the problem domain, and a series of breakout sessions, where the participants discussed specific exercises more detailled in three groups. The groups were composed of a selected set of researchers, ensuring a good mixture of seniority and juniority, as well as a coverage of all the aspects required to address the problems at hand. The breakouts were designed as competitions between the group, whose results were evaluated in the follow-up plenary sessions.

The three group leaders were:

- David Lowenthal
- Frank Mueller
- Martin Schulz

The experience with these breakouts exceeded expectations by leading to new results and also extensive contributions to the discussions by all participants. As such, the structure of this workshop proved very useful and might be a good idea for other topics as well.

5 Open Problems – Future Research Direction

The workshop identified a series of major problems, each covering a number of technical issues as future research topics. The major problems are as follows:

- *Different groups have different optimization functions*: While the overall goal, efficient usage of power at the highest level of performance, is the central goal, the actual goals for each group depend on their respective layer in the computing stack. We identified different goals for the layers infrastructure, system software, algorithms, and applications. In addition, the goals may also differ between the computing centers corresponding to their respective targets and operational goals.

- *Information exchange between layers is insufficient*: In order to achieve optimal performance in a power-bound environment, improved information exchange between the above mentioned layers is needed. This requires corresponding tools and interfaces, such that the information available on one layer can be transferred to the layers above or below.
- *User happiness must be weighted against total costs against performance*: While solutions for one characteristic are possible, they have to be weighted against each other. Any solution needs to ensure that application users are happy enough with the operation of the machine, while providers are able to shoulder the costs, while the performance of the application offers a suitable time-to-solution.

Participants

- Michael Bader
TU München, DE
- Natalie Bates
Lawrence Livermore National Laboratory, US
- Pete Beckman
Argonne National Laboratory, US
- Jonathan Eastep
Intel – Hillsboro, US
- Neha Gholkar
North Carolina State University – Raleigh, US
- Joseph Greathouse
AMD – Austin, US
- Thomas Ilsche
TU Dresden, DE
- Dieter Kranzlmüller
LMU München, DE
- Stephanie Labasan
Univ. of Oregon – Eugene, US
- David K. Lowenthal
Univ. of Arizona – Tucson, US
- Matthias Maiterth
LMU München, DE
- Andres Marquez
Pacific Northwest National Lab. – Richland, US
- Yousri Mhedheb
KIT – Karlsruher Institut für Technologie, DE
- Shirley V. Moore
University of Texas – El Paso, US
- Frank Mueller
North Carolina State University – Raleigh, US
- Andreas Raabe
DFG – Bonn, DE
- David Richards
LLNL – Livermore, US
- Suzanne Rivoire
Sonoma State University – Rohnert Park, US
- Barry L. Rountree
LLNL – Livermore, US
- Martin Schulz
LLNL – Livermore, US
- Kathleen Sumiko Shoga
LLNL – Livermore, US
- Torsten Wilde
LRZ – München, DE

Report from Dagstuhl Seminar 15351

Computational Mass Spectrometry

Edited by

Rudolf Aebersold[1], Oliver Kohlbacher[2], and Olga Vitek[3]

1 **ETH Zürich, CH,** aebersold@imsb.biol.ethz.ch
2 **University of Tübingen and Max Planck Institute for Developmental Biology, DE,** oliver.kohlbacher@uni-tuebingen.de
3 **Northeastern University, US,** o.vitek@neu.edu

Abstract

Following in the steps of high-throughput sequencing, mass spectrometry (MS) has become established as a key analytical technique for large-scale studies of complex biological mixtures. MS-based experiments generate datasets of increasing complexity and size, and the rate of production of these datasets has exceeded Moore's law. In recent years we have witnessed the growth of computational approaches to coping with this data deluge.

The seminar 'Computational Mass Spectrometry' brought together mass spectrometrists, statisticians, computer scientists and biologists to discuss where the next set of computational and statistical challenges lie. The participants discussed emerging areas of research such as how to investigate questions in systems biology with the design and analysis of datasets both large in memory usage and number of features and include measurements from multiple 'omics technologies.

Seminar August 23–28, 2015 – http://www.dagstuhl.de/15351
1998 ACM Subject Classification J.3 Life and Medical Science
Keywords and phrases computational mass spectrometry, proteomics, metabolomics, bioinformatics
Digital Object Identifier 10.4230/DagRep.5.8.9
Edited in cooperation with Robert Ness and Timo Sachsenberg

1 Executive Summary

Robert Ness
Timo Sachsenberg
Rudolf Aebersold
Oliver Kohlbacher
Olga Vitek

Motivation

Mass Spectrometry (MS) is an extremely flexible analytical technique, with applications ranging from crime lab investigations to testing to disease biomarkers in a clinic. The publication of the first human genome in 2001 was a key event that lead to the application of mass spectrometry to map out the human proteome, and later the human metabolome; i.e. all the biomolecules encoded in the genome that constitute biological function. The result was the creation of a tremendous amount of spectrometric data and a dearth of tools for data

Computational Mass Spectrometry, *Dagstuhl Reports*, Vol. 5, Issue 8, pp. 9–33
Editors: Rudolf Aebersold, Oliver Kohlbacher, and Olga Vitek
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

analysis, motivating the development of computational tools. The tool developers came from several expert domains; life scientists applying mass spectrometry built tools to automate their new workflows, analytical chemists and engineers developing the instruments built software to analyze devise measurements; network and database infrastructure professionals built resources for storing and sharing data in the cloud, and bioinformaticians and statisticians developed algorithms and statistical methods for data analysis. There is an ongoing need for the different disciplines to learn each other's languages, make tools interoperable, and establish common goals for development.

Goals

The seminar 'Computational Mass Spectrometry' is a follow-up seminar to the successful Dagstuhl seminars on 'Computational Proteomics' and 'Computational Mass Spectrometry' (05471, 08101 and 14371).

The seminar aimed at bringing together scientists from a wide range of backgrounds and identify open issues and future research directions in computational mass spectrometry.

Results

Already on the first days the seminar resulted in very lively discussions. The time allotted to the introductory talks had to be expanded to account for this. The discussions sparked off during the introductory talks led to the formation of several working groups. These groups formed and re-formed on demand, also based on discussion on the previous evenings. Section 5 documents the discussions and results in these groups through the notes taken. Some of these discussion (e.g., the one on false discovery rates) was of interest to all participants and took place as plenary discussions in the large lecture hall. Other discussions were more focussed and thus had a smaller number of participants.

Some of the discussion will certainly lead to joint research participants. A first tangible outcome is a joint paper already accepted in the *Journal of Proteome Research* (L. Gatto, K. D. Hansen, M. R. Hoopmann, H. Hermjakob, O. Kohlbacher, A. Beyer, "*Testing and validation of computational methods for mass spectrometry*," DOI: 10.1021/acs.jproteome.5b00852) on benchmarking and validating computational methods for mass spectrometry. This working group developed conceptual ideas for benchmarking algorithms and implemented a web-based repository holding (http://compms.org/RefData) benchmark datasets that will hopefully make comparison of algorithms more transparent in the future. We are confident that the discussions of other working groups and the contacts made during the evening hours in Dagstuhl will result in many more collaborations and publications in the future.

The field of computational mass spectrometry is rapidly evolving. Participants identified a wide range of challenges arising from technological developments already at the horizon but also from the broadening on the application side. We thus intend to revisit the field in the coming years in a Dagstuhl seminar again, most likely organized by different leaders of the field in order to account for these upcoming changes.

2 Table of Contents

3 Structure of the Seminar

The seminar was structured into introductory talks by participants from diverse fields of mass spectrometry. After the overview talks, proposals for break-out group topics were collected. These were aimed at allowing for more focused discussions in smaller groups. The participants then voted on these topics. Work groups (WG) were formed every morning over the whole course of the Dagstuhl seminar. Overview talks were limited to the first two days and had been solicited by the organizers well in advance. Teams of two to three participants were given the task to present a topic they are experts in with the purpose of introducing the other participants to the field as well as getting a personal view on the state of the field.

The first two days of the Dagstuhl seminar was intended to give a broad overview of current topics in computational mass spectrometry with a focus on the challenges of dealing with large data, common misconception of statistical problems associated with their analysis as well as the integration of data of different omics technologies. The remaining days intensified the discussion on central aspects of these challenges in break-out groups. We were very happy to include the seminar on microfluidics (which was held in parallel at Dagstuhl) into a joint morning session on Wednesdays.

The overall schedule of the seminar was as follows:

Monday

- Welcome and introduction of participants
- Computational mass spectrometry – the big picture (introductory talk)
- Challenges in metabolomics
- Statistical methods

Tuesday

- Reproducibility and big (omics) data
- Democratization of omics data
- Multi-omics data integration
- Spatial aspects of multi-omics
- System dynamics based on multi-omics data

Wednesday

- Joint session with Dagstuhl Seminar 15352 “Design of Microfluidic Biochips”
- Breakout groups
  1. WG ‘Big Data & repositories’
  2. WG ‘Correlation vs. causality’
  3. WG ‘Testing and validation of computational methods’
  4. Outing: World Cultural Heritage Site Völklingen Ironworks

Thursday

- Joint session: reports on the Wednesday sessions
- Break-out groups
  1. WG ‘Multi-omics case studies’
  2. WG ‘Metabolomics and proteomics integration’
  3. WG ‘Systems genetics’

Friday

- Breakout groups
  1. WG ‘Metaproteomics’
  2. WG ‘Computational challenges in quantitative proteomics’
  3. WG ‘Validation and Reference datasets’
  4. WG ‘Education’
  5. Seminar wrap-up and departure

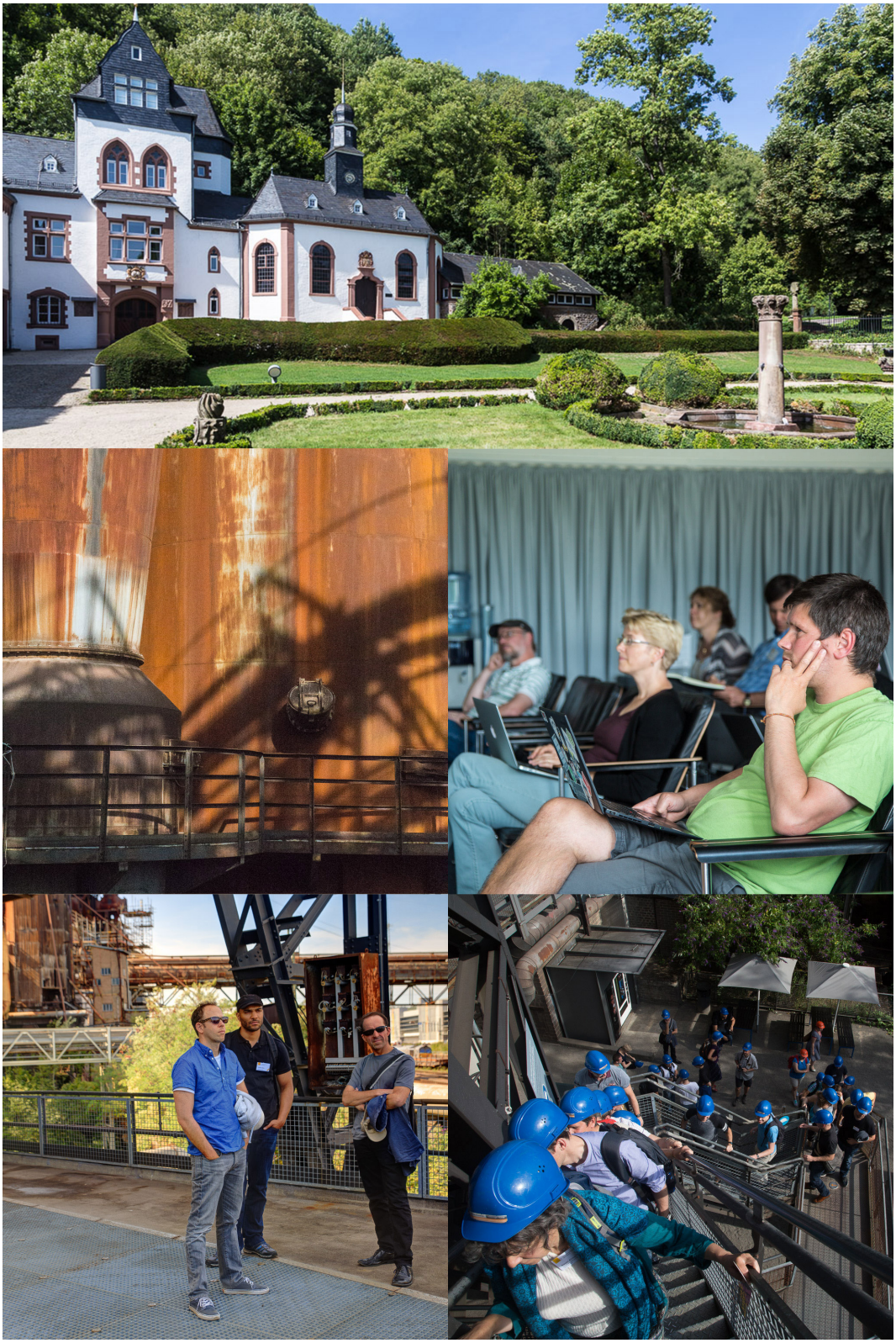


Figure 1 Some impressions from the seminar and the outing at Völklingen ironworks (photos: Oliver Kohlbacher, Pedro Navarro).

4 Overview of Talks

4.1 Challenges in Computational Mass Spectrometry – Objectives and Data Collection

Rudolf Aebersold, ETH Zürich, CH

The proteome catalyzes and controls the ensemble of essentially all biochemical reactions of the cell and its analysis is therefore critical for basic and translational biology. The proteome is also exceedingly complex with potentially millions of different proteoforms being expressed in a typical mammalian cell. In this presentation we will discuss and assess the current state of mass spectrometric methods to identify and quantify the components of the proteome with two primary objectives. The first objective is the generation of a complete proteome map of a species, i.e. a database that contains experimental evidence for every protein or proteoform expressable by a species. The second objective is the generation of large numbers of highly reproducible, quantitative proteome datasets that represent different states of cells and tissues to support the study of the dynamic adaptation of biological systems to perturbations.

4.2 Challenges in Computational Mass Spectrometry – Statistics

Olga Vitek, Northeastern University – Boston, US

'Big data' has passed it's 'hype' point, and it is now time to enter a 'productivity stage. Statistical methods are key for this task. They need to address several challenges, for example; (1) larger datasets can hide small signals, (2) give rise to spurious associations, (3) encourage researchers to mistake association for causality, and (4) give rise to bias and confounding. The fundamental principles of statistical design and analysis, and domain knowledge, are key for avoiding these pitfalls.

4.3 Challenges in Computational Mass Spectrometry – Data and Tools

Oliver Kohlbacher, Universität Tübingen, DE

Computational mass spectrometry currently faces several challenges from the ever growing volume and complexity of the data. This is caused by the increase in instrument resolution and speed, new acquisition techniques, but also by the need for parallel application of several high-throughput methods in parallel (multi-omics). Lack of interoperability and usability of bioinformatics tools currently hampers the analysis of large-scale data and has also implications for reproducibility – and thus the reputation – of MS-based omics techniques.

4.4 Spatial Metabolomics: Why, How, and Challenges

Theodore Alexandrov, EMBL Heidelberg, DE

Spatial metabolomics is emerging as a powerful approach to localize hundreds of metabolites directly from sections of biological samples with the grand challenge to be in the molecular annotation of big data generated. We will present Why spatial metabolomics may be important, How it can be performed and overview computational Challenges. Computational Mass Spectrometry is essential in this field, since existing bioinformatics tools cannot be applied directly because of the sheer data size and high complexity of spectra. We will also present algorithms for molecular annotation for High Resolution Imaging Mass Spectrometry that integrates both spectral and spatial filters. We will present the European project METASPACE on Bioinformatics for Spatial Metabolomics.

4.5 Some Statistical Musings

Naomi Altman, Pennsylvania State University – University Park, US

Musings on a set of statistical topics that might be interesting in MS studies:

- feature matching across samples and platforms
- preprocessing and its effects on multi-omics
- analysis problems when the number of features is larger than the number of samples
- feature screening
- replication and possibly other design issues
- dimension reduction via PCA and related methods
- mixture modeling

4.6 Reproducibility and Big (Omics) Data

Nuno Bandeira, University of California – San Diego, US
Henning Hermjakob, European Bioinformatics Institute – Cambridge, GB

The volume of omics data, including mass spectrometry-based proteomics, approximately doubles every 12 months. At EMBL-EBI, mass spectrometry data is now the second largest data type after sequence data. In the last three years, the ProteomeXchange consortium has established a collaboration of databases to ensure efficient and safe provision of data to the community, currently processing more than 200 submissions per month, and supporting a download volume of 150+ TB/year. Strategies for data access comprise cloud-based processing of raw data, common APIs for data access across multiple resources, and a transition from static data submissions to dynamic re-analysis of data in the light of new computational approaches and database content. Beyond data size and complexity, Proteomics now has to

face the challenge of personally identifiable data, as the resolution of proteomics methods now allows to associate a proteomics dataset with its source genome due to identification of amino acid variants.

4.7 Introduction to Metabolite Mass Spectrometry

Sebastian Böcker, Universität Jena, DE
David Wishart, University of Alberta – Edmonton, CA

Metabolites, small molecules that are involved in cellular reactions, provide a direct functional signature of cellular state. There is a large overlap between metabolomics and proteomics with regards to the experimental platform used for high-throughput screening, namely, mass spectrometry and tandem MS. In our talk, we have highlighted both similarities and differences between the fields.

A particular noteworthy difference between the fields is that the identification of a peptide via tandem MS is a somewhat straightforward problem, whereas the same is highly non-trivial for metabolite ID. We discussed reasons for this, in particular the structural diversity of metabolites, and our inability to predict a tandem MS for a given metabolite structure. We then discussed approaches to overcome this problem: namely, combinatorial fragmenters (MetFrag, MAGMa), prediction of spectra using Machine Learning and MCMC (CFM-ID), and the prediction of molecular fingerprints from tandem MS data ((CSI:)FingerID).

4.8 Democratization of Data: Access and Review

Robert Chalkley, University of California – San Francisco, US

Studies that are published in a peer-reviewed journal are supposed to come with a guarantee of reliability. For large omics studies a reviewer cannot be expected to re-analyze data, so there is a need for the community as a whole to evaluate data and results. This places a high pressure on journals to capture sufficient meta-information about data and analysis to permit appropriate reanalysis. This presentation describes the current status of publication guidelines of the journal Molecular and Cellular Proteomics, as a representative of publishers in this field. It also provides a discussion of the blurring line between a journal publication and a submission of data and results to a public repository, which also requires provision of certain metadata.

4.9 Multi-omics Data Integration

Joshua Elias, Stanford University, US

As high throughput technologies for measuring biological molecules continue to improve, so will researchers' need to combine them. Each domain of such 'omic' technologies has a distinctive set of pitfalls that may not be readily apparent to non-experts: Techniques focused on nucleic acids (genomics, transcriptomics, metagenomics, translatomics), proteins (proteomics) and metabolites (metabolomics, lipidomics, glycomics) range widely in several important features: Instrumentation required for reliable measurements; methods for evaluating measurement error, quantitation accuracy and precision, data format, and visualization tools. As a result, experts within individual domains and often sub-domains need to cooperate in order for large, multi-omic experiments to be carried out successfully. Major challenges and opportunities exist for improving analytical standards within omic domains such that their results can be directly aligned, and confidently assimilated for interdisciplinary research.

4.10 Some lessons from Gene Expression

Kasper Daniel Hansen, Johns Hopkins University – Baltimore, US

We discuss statistical lessons learned from the analysis of gene expression data, including experimental design, batch effects, reproducibility and data availability.

4.11 Spatial Proteomics

Kathryn Lilley, University of Cambridge, GB

Cells are not just collections of proteins randomly distributed in space. Proteins exist in restricted sub-cellular niches where they have access to substrates/binding partners/appropriate chemical environments. Many proteins can exist in multiple locations and may adopt different roles in a context specific manner. Sampling the spatial proteome is non trivial. Moreover proteins redistribution upon perturbation may be as important feature to capture as change in abundance or post translational status. There are multiple methods to capture the spatial proteome. Some of these are based on existing hypotheses, where the proteome is tested on a protein by protein basis per experiment, for example immunocytochemistry approaches. Other methods capture the 'local' proximity of proteins by directed labelling of surrounding proteins to the protein of interest and downstream analysis of the labelled entities. Developing approaches attempt to establish the steady distribution of proteins within sub-cellular niches on a cell-wide scale.

The emerging methods are highly complementary, but all are associated with technical and analytical challenges. The different broad approaches and their specific challenges are discussed in this presentation.

4.12 Democratizing Proteomics Data

Lennart Martens, Ghent University, BE

A view on democratizing data, with emphasis on local data management and a path from quality control to accreditation.

4.13 System Dynamics from Multi-Omics Data

Karen Sachs, Stanford University, US

Given sufficient data, it is possible to extract network regulatory information from multi-dimensional datasets. I will first present a short tutorial on probabilistic graphical modeling applied to network inference, using the example of single cell proteomics data. Next, I'll discuss the impact of time and our ability to extract dynamic models from these data.

4.14 Considerations for Large-Scale Analyses

Michael L. Tress, CNIO – Madrid, ES

We interrogated a conservative reliable set of peptides from a number of large-scale resources and identified at least two peptides for 12,000 genes. We found that standard proteomics studies find peptides for genes from the oldest families, while there were very few peptides for genes that appeared in the primate lineage and for genes without protein-like characteristics.

We found similar results for alternatively spliced exons – we found few, but those we did find were of ancient origin. The sixty homologous exon splicing events we detected could be traced all the way back to jawed vertebrates, 460 millions years ago.

Our results suggest that large-scale experiments should be designed with more care and those that identify large numbers of non-conserved novel coding regions and alternative splice events are probably detecting many false positives cases.

4.15 System Dynamics Based on Multi-Omics Data

Nicola Zamboni, ETH Zürich, CH

The current standards of transcriptomics, proteomics, metabolomics, etc. allow to simultaneously profile/quantify large number of molecules in cellular systems and biofluids. In the field of cell biology, comparative analysis of two or more groups often results in discovering a

multitude of statistically significant differences. Such complex patterns result from the overlap of primary and secondary effects caused by cellular regulation and response. Translation of such results into testable hypotheses suffers from two fundamental problems. First, human intuition doesn't scale enough to integrate several changes in the context of large metabolic networks. Second, analytical methods allow us only to assess changes in composition (state), but not on the integrated operation (activity). Hence, omics data provide only an indirect readout that we can't simply associate to a functional change. This calls for computational methods that infer testable hypotheses on the basis of omics information and previously known networks. Such approaches can be supported experimentally by (i) performing time-resolved experiments with multiple datapoints, or (ii) generation of reference datasets in which the omics profile has been recorded for known perturbations under comparable conditions.

5 Results from the Working Groups

Working groups were formed and re-formed throughout the whole seminar. At the beginning of each day, groups reported on their results. Some topics attracted the interest of the whole audience and were selected for joint sessions. Other more specialized topics led to formation of medium or small groups.

5.1 Big Data and Repositories

Susan Weintraub, Lennart Martens, Henning Hermjakob, Nuno Bandeira, Anne-Claude Gingras, Bernhard Kuster, Sven Nahnsen, Timo Sachsenberg, Pedro Navarro, Robert Chalkley, Josh Elias, Bernhard Renard, Steve Tate, and Theodore Alexandrov

The group mostly focused on the question of the interactions between the mass spectrometry repositories and the scientific community. Interactions are with publishers / reviewers, data providers, computational tool developers, "end-user" biologists, etc. All participants agreed that repositories are important, and that much of the minutiae of data standards and repository organization have already been sorted out. Therefore, the discussion mostly centered on the design of useful features for the community using the data in the repositories. While repositories have worked in the past in a linear manner where the data depositor (user; U), after employing tools developed by software designers (S) would submit their data in the repositories. On the part of the user, one of the biggest incentive was to fulfill the requirements for publications. However, now that the repositories are up and running, the data depositor could be further incentivized by having the repositories providing additional value to their data.

Journal deposition requirements. How to best support the publication/validation process? Some way to support the process include; (1) automatic generation of a methods section summary with aggregate results views (e.g., FDR/ROC curves, LC-MS thumbnail , run-to-run or condition-to-condition comparison), (2) Ability to search for spectra (file name + scan), (3)

derive new knowledge reprocessing guidelines "dataset notes" type of manuscripts, (4) having living datasets for "ongoing iPRG" benchmarking. A key problem concerns metadata. For example, submissions typically fail to include aquisition parameters in metadata. More general metadata questions include; what should be required, what should be merely recommended or altogether discarded? How should we distinguish technical from biological replicates? Other avenues for improvement include e-mail or detail views (e.g. for reviewers). One issue with multi-omics submissions is the size of the data. How to compute on big data? Should we invest in big data analysis tools within our repositories? Medical/Clinical data cannot (easily) move to public clouds for either private compute or repository access.

Algorithmic challenges. APIs Bringing tools to the data? What views should repositories aim to provide to a) biologists, b) biostatisticians, c) bioinformaticians, d) other?

Data repositories from a biologist's perspective. Biologists want: peptide and protein expression levels across datasets and conditions. What incentives/benefits to provide to data submitters? How to add value to the data (e.g., like genome browser)? Cover as many instruments as possible. Spectrum clusteringto find most similar datasets. Protein view with peptide coverage and detected PTMs. Ability to link peptides to spectrum data. Match my search results against repository. Protein coverage.

Dataset-centric view. Which proteins/peptides/PTMs/sites does it contribute the most to? Which proteins/peptides/PTMs/sites is the dataset missing that it should be seeing? Links to other repositories: CRAPome, UniProt, ProteomicsDB, PDB, Protein Atlas. Sync protein identifiers to cross-reference to AP/interactions repositories. Cross-reference peptides by sequence Repository APIs for cross-references reference data: Bernhard Küster offered deposition of synthetic peptide spectra.

Quantitative views. ProteomicsDB gene/protein list linked to expression levels across datasets. Download as table, filter by type of quant (e.g., SILAC, TMT); Label-free is less biased to experiment design.

5.2 Integration of Metabolomics and Proteomics

Jonathan O'Brien, Nicola Zamboni, Sebastian Böcker, Knut Reinert, Timo Sachsenberg, Theodore Alexandrov, Henning Hermjakob, and David Wishart

5.2.1 State of the Art

General Comment. Despite sharing similar instrumentation and relatively similar computational needs there is relatively little integration between the two fields. We discussed some existing and emerging examples of where the two fields have connected or could interact.

Existing Examples. One example of proteomic/metabolomics integration has been through systems biology studies involving the characterization of cells (yeast, E. coli) and humans through combined experimental and computational efforts (Human Recon2, Yeast Metabolic Reconstruction, IFBA). These have led to computational constructs that model metabolite fluxes and flows and which could predict certain phenotypes or diseases based on mutations, knockdowns or knockouts of genes and proteins. This work led to the development of

SBML and the development flux-balance models, ODEs, petri-nets, PDEs and agent-based models for cell simulation. However, the SB field struggles because the omics data is often incomplete and insufficiently quantitative to go beyond "toy" models. Another example of integration has been the creation of pathway databases that depict protein and metabolite data with qualitative indications of abundance or presence/absence. Examples include KEGG, Cyc-databases, Reactome, Wikipathways, SMPDB. However, the model needs and mark-up languages used by the metabolomics community (KEGG-ML, SBML, PGML) are often incompatible with the model needs or mark-up languages used by the systems biology and proteomics community (SBGN-ML, BioPax)

Emerging Examples. An emerging area of experimental proteomics that integrates metabolomics with proteomics is called Adductomics, which is part of the field of Exposomics. This measures the chemical modifications of electrophilic adducts to free cysteines in serum albumin or other groups in hemoglobin. This is used to detect and quantify the presence of pollutants, toxins and reactive drug byproducts in organisms. Currently the field of adductomics lacks software tools and databases to facilitate the characterization of the peptides and products. Another emerging area of experimental proteomics that impacts metabolomics is MS-based protein-ligand screening and MS-based binding constant measurement. Normally this is used in drug discovery but potentially this could be used to rapidly screen which proteins bind to which metabolites (proteome-to-metabolome-mapping). However, this field lacks software tools and databases to do this rapidly and efficiently.

What can proteomics learn from metabolomics and vice versa?

1. A major focus of proteomics is on deciphering signaling networks while the major focus on metabolomics is describing catabolism and anabolism. The result is the proteins are viewed as "brains" in the cell while metabolites are just the bricks and mortar. Most software tools and databases in proteomics focus on protein signaling, but most software tools in metabolomics focus on anabolism and catabolism. The interpretation of metabolomics data needs to include metabolite signaling. We've forgotten that the primary role of metabolites is actually to signal proteins. A problem is that none of the metabolomic databases have this information. However, some proteomics databases (Reactome, Wikipathways, SMPDB) do – but not enough of it or not in a useable form. Action item: The metabolomics community needs to learn from the proteomics community and think about deciphering signaling pathways, too. Metabolite signaling data is available in books, journals and on-line protein-pathway databases, but it is not machine readable or not compatible with current versions of metabolomics software or current needs of metabolomics researchers. There is a clear gap between the communities and community standards – the two communities need to work together to get this sorted out. It is proposed that representatives of the metabolomics community attend the next COMBINE meeting[1] (SBML/BioPAX/SBGN-ML standards meeting).
2. A major focus of metabolomics is targeted, quantitative studies where small numbers of metabolites are measured with absolute concentrations. In contrast in proteomics, the focus is measuring large numbers of proteins with relative or semi-relative concentrations. Because metabolomics is becoming more quantitative it is allowing computational scientists to work on biomarker identification and allowing them to mine existing data to discover new biomarkers and biomarker combinations. It's also allowing metabolite discoveries to transition to clinical applications quite quickly. There are now >160

[1] http://co.mbine.org/

metabolite tests used in the clinic. More than a dozen quantitative metabolomics kits are now commercially available and easy/cheap to run. Quantitative data also allows researchers to compare data sets across labs or studies and to perform meta-data analysis more consistently. However, proteomics still lags behind other fields in its ability to quantify (absolutely or qualitatively) Action item: The proteomics community needs to learn from the metabolomics community and think about ways of generating (via kits?) and archiving targeted (or non-targeted) quantitative proteomics data. The use of common data storage formats and common experimental description formats would help. Specifically mzML, mzTAB and mzQuantML need to be used and adopted by both communities. Agreement on how to quote or measure protein concentration data (in absolute terms) would help. It is proposed that representatives of the metabolomics community attend the next mzML, mzTAB and mzQuantML standards meeting (PSI Spring meeting 2016 in Gent, Belgium).

3. Proteomics has evolved a much more sophisticated system for quality control at the instrument and data collection level (OpenMS). Metabolomics has evolved very sophisticated systems for quality control at the sample handling and sample comparison level (MetaboAnalyst). However, the metabolomics community is not utilizing the mzTAB format while neither community is utilizing the mzQuantML sufficiently. Action item: The two fields should borrow the tools that the others have developed so that both can improve QC at both the instrument and sample handling levels. Both need to make better use of existing data standards and data exchange formats
4. Genomics measures or sequences genes at an "organism level", Metabolomics tends to measure fluids at the "organ level" while proteomics and transcriptomics measures protein/gene abundance at a cell or "tissue level". This can make integration difficult and comparisons challenging. Action item: More discussion needs to be had about how the fields can come to a more common unit of measurement. Should proteomics focus more studies on biofluids? Should metabolomics focus more on studying tissues? Should proteomics and metabolomics be done simultaneously on the same sample?

Open Questions

1. Can we go beyond mapping quantities to pathways? What about including dynamics? How to include or measure transient protein-metabolite interactions? What about complexes (metabolites and proteins)?
2. Can we get the 2 communities talking together on a more regular basis? (bioinformaticians, standards and focused meetings are key)
3. Primary metabolism in good state but many difficulties with promiscuous enzymes (might be bridges to complete network) but not secondary metabolism – we are missing most of the proteins, interactions and pathways for these processes. What to do?
4. How to deal with the problem of relative quantification vs. absolute quantification?
5. How do the two communities handle issues of pathway plasticity?
6. Is proteo-metabolomics possible? Can the combined data be loaded into an appropriate repository anywhere?
7. Can metabolomics be used to better characterize the phenotype to help "amplify" the proteomic trends or proteomic findings?

5.3 Multi-Omics Case Studies

Pedro Jose Navarro Alvarez, Joshua Elias, Laurent Gatto, Olga Vitek, Kathryn, Karen Sachs, Rdolf Aebersold, Oliver Kohlbacher, Stephen Tate, and Christine Vogel

This area seems to follow the same pattern as many hyped fields: excitement, confusion, disillusion, and realism.

Excitement. First studies available – see case studies above: integrating proteomics and transcriptomics data at steady state or from time series experiments, complemented by ribo seq data also: papers such as Aviv Regev's (Science 2015).

Confusion. What do correlations mean? What do we learn from them? [Olga, Christine] We need more complex approaches, e.g. dynamic models. [Oliver] But there are many dynamic models. It depends on your question what you need to do.

Disillusion. [Oliver] Do we have a common language for data integration? – [Kathryn] Do we need one? How do we get started on integrating different errors/noise estimates, FDRs, data types? So much noise, so much complexity to the data, so many different error models, so different data structures – where do we start? Where do we start if the data type we understand best (proteomics data) already has big problems?

Realism. What do we actually mean by integration? Be clear about your biological question (as usual). [Ruedi] Even simple models illustrate that we do not really know how biology works. Even in proteomics, the domain we know most about, it is difficult to make meaningful predictions. How do we take the omics data with limited knowledge behind it and use it in a useful way and learn something new? Go slow: carefully consider your data and its properties. Use smaller, well-defined systems. E.g. [Karen's example] [lunch discussion]. Don't forget your biology (or biologist). Stare at the data (and don't ignore odd things). Use the scientific method: generate hypotheses based on your data and test them. Do we need integrative tools? Is it time already? [Oliver] Yes e.g. Perseus is moving towards that – PRIDE as well? e.g. use RNA to help identification of peptides in MS data (proteogenomics).

5.4 Testing and validation of computational methods

Participants: Andreas Beyer, Kasper Daniel Hansen, Laurent Gatto, Michael Hoopmann, and Oliver Kohlbacher

The goal of this group was to discuss means for testing, validating, and comparing computational methods, focusing – of course – on methods dealing with proteomics data. It is perhaps trivial to identify bad computational methods, but more difficult to recognize the best methods. We did not distinguish statistical and computational methods, but we distinguished experimental method validation from computational method validation. The discussion mostly dealt with methods for peptide identification and protein level quantification, but we feel that the conclusions are much more widely applicable. Further, we emphasized that the

way how methods be validated will depend a lot on the specific problem, e.g. the difference between absolute protein quantification versus quantification of fold-changes. Hence, it is crucial to identify and documente measurable outcomes (objective metrics) underlying the comparison.

1. **Too many user-definable parameters.** Usable computational methods should not have too many user-definable parameters. Methods with many parameters cause two problems: (1) It becomes difficult for end-users to correctly set the parameters and experience shows that for real-life applications most people will use default settings. (2) Comparing methods becomes exceedingly difficult if many possible combinations of parameters have to be tested and compared against other methods. Further, having many parameters creates the danger that users might tweak parameters until they get a 'desired' result, such as maximizing the number of differentially expressed proteins. We therefore came up with the following recommendations: Methods should have as few user-definable parameters as possible. If possible, parameters should be learned from the data (e.g. via built-in cross validation.) If user-definable parameters are unavoidable there should be very clear instructions on how to set these parameters depending on the experimental setup. (E.g. depending on the machine used, species the samples come from, goal of the experiment, ...)
2. **Simulated data.** A risk of using simulated data is that the simulation will reflect the implicit model underlying a computational method. There is a continuum to the extend simulated data will reflect reality. Reliance and wide acceptance of simulation might be reached using community-accepted simulator, rather than project-specific driven simulations. We however recognise some value to simulation, to understand method and a sophisticated code checking mechanism, and understand effects, stability of methods rather than compare them. Comparisons based on simulations should be interpreted with care and complemented by utilization of real data (see below).
3. **Reference data, spike in data, etc.** Spike-in should be sufficiently complex to thoroughly challenge methods (e.g. spike into a 'real' sample). Negative controls need to be included (e.g. known static proteins in data mixed with proteins changing quantity). Gold-standard sets are important, but can lead to biases the optimize against the gold-standard. More than one reference set should be tested. Reference sets need not be immaculate data.
4. **Use of real data, multi-omics.** We identified an opportunity to initiate a debate on multi-omics reference datasets to support methods development and comparison. Using real data without a well-defined 'ground truth' requires creativity, but it is not impossible. Importantly, external, independent data can be used as a common reference to compare outputs of different analysis methods to. For example, expect that protein concentrations should be somewhat correlated to their mRNA concentrations. Thus, protein and mRNA data coming from identical samples could be used to evaluate the performance of different protein quantification methods: if one method results in significantly greater correlation between protein and mRNA than another, that could be used as a guideline for choosing the method. We agreed that such data sets could be very valuable and should be made available to the community. These thoughts sparked a general discussion around the opportunities of combining multi-omics data from matching samples. We expect a great potential of such analyses also for improving computational methods.
5. **Community resource for reference datasets.** We concluded that the community would benefit from a resource with guidelines, suggestions, references, ... summarising the above reflection, that we would like to initiate. We will reach out to the seminar delegates and

the community for material for method development and comparison, such as reference data sets (for example spiked-in data), data simulators, useful papers and methods.

Reference data

- Benchmark datasets for 3D MALDI- and DESI-imaging mass spectrometry:
  http://www.gigasciencejournal.com/content/4/1/20
- Data for comparison of metabolite quantification methods (including spike-in datasets and simulated datasets):
  http://www.mcponline.org/content/13/1/348.long
- Protein Identification and inference benchmarking dataset:
  http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.5b00121
  Corresponding datasets are in PRIDE (PXD000790-793)
- Validation data set for functional metaproteomics based on stable isotope incorporation:
  http://pubs.acs.org/doi/abs/10.1021/pr500245w (PRIDE PXD000382)
- A published DIA/SG data set comprising 8 samples with stable HEK-293 background and several proteins spiked in in different known absolute amounts. The spike in differences are small changes, large changes and span a large dynamic range. The 8 samples were measured in triplicates and in DIA and shotgun (48 measurements) on a QExactive. We used the data set to compare the quantitative hallmarks between DIA/SG, i.e. missing values, CVs and accurate of fold change detection. The data set can be used to benchmark quantitation, algorithms for DIA analysis and probably other things.
  https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View?identifier=PASS00589
  and publication http://www.mcponline.org/cgi/pmidlookup?view=long&pmid=25724911
- The ABRF iPRG 2009 for label-free differentiation:
  ftp://massive.ucsd.edu/MSV000078539
- For PTM discovery, the FFPE tissues:
  ftp://massive.ucsd.edu/MSV000078985
  CPTAC provides a standard dataset (Study 6) in which Sigma UPS1 (48 equimolar proteins) are spiked into yeast at different dilution factors. The sample is analyzed by shotgun MS using HPLC+ESI. The dataset can be found at:
  https://cptac-data-portal.georgetown.edu/cptac/study/list?scope=Phase+I
  The dataset has been analyzed on multiple instruments for added versatility. I consider the quality as medium. Several publications describing the dataset and analyses performed are found at:
  http://www.ncbi.nlm.nih.gov/pubmed/19858499
  http://www.ncbi.nlm.nih.gov/pubmed/19837981 and
  http://www.ncbi.nlm.nih.gov/pubmed/19921851
  PXD001500 is excellent for quantitative MudPit, to be testes for carbamylation at K and nt PXD001792 is excellent survey phosphorylation data PXD002140 is excellent prokaryote survey data
- Simulators:
  http://www.ncbi.nlm.nih.gov/pubmed/25371478
  http://www.ncbi.nlm.nih.gov/pubmed/24090032
  http://www.ncbi.nlm.nih.gov/pubmed/21526843
  http://www.biomedcentral.com/1471-2105/9/423

5.5 Systems genetics

Andreas Beyer, Hannes Röst, Matthias Gstaiger, Lukas Käll, Bernard Rennard, Kasper Hansen, Stefan Tenzer, and Anne-Claude Gingras

We primarily discussed complex diseases. For monogenic disease, omics and proteomics in particular can be very useful in defining the mechanism underlying disease, but here we primarily focused on complex diseases, or complex genotype-phenotype relationships. Typically this would be taking some kind of genetic analysis, such as GWAS, or QTLs, or cancer mutations. Then we would use omics tools (multi-omics, though proteomics and transcriptomics were mostly discussed) to provide a better view of genotype-to-phenotype relationships. Why multi-omics? The potential benefits of multi-omics in this context were at least twofold: (1) Improving the identification of causing mutation and (2) improving the understanding of the molecular mechanisms.

- **Improved identification of causal variants.** Conceptually, Omics data can improve genetic mapping in two ways: GWAS/QTL datasets with multiple genes in an identified locus may be better teased apart (e.g. protein levels can help with the fine mapping of the causal gene/protein) Multi-omics can bring increased sensitivity. Statistically weak GWAS associations may not be found without omics data. For example, network analysis, SNP clustering, etc. may help better interpreting the data.
- **Revealing molecular mechanisms.** For understanding the molecular mechanisms, at the simplest level, one can consider many multiple omics (particularly expression omics) as a massively multiplexed phenotypical readout of the effect of the perturbed genome. Mutations could impact the transcriptional or post-transcriptional regulation of gene expression. This is the first manifestation of these mutations. An example is a mutation in a transcriptional regulator that would generate a molecular fingerprint of its transcriptional targets. Conversely, a kinase could potentially be identified by profiling the phosphoproteome. Expression proteomics are important to uncover regulation, e.g. of protein stability, that would not be uncovered by profiling RNA expression alone. To get at the molecular mechanism underlying these changes, other omics technologies can also be used. Differential interaction proteomics are particularly useful, but require pre-filtering since they do not scale well to the growing list of genetic alterations.
  Types of omics-data integration: There is a distinction to be made between overlapping datasets and integrating datasets, both of which being useful. This is a continuous scale. Overlapping datasets involve completely separate analysis of each omics technology results and then comparing the results. There is no information feedback between omics technologies. Integrating datasets entails simultaneous analysis of both datasets. In some cases, one omics / analysis improves the analysis of the other. Alternatively, you can extract new information from integrating both datasets that could not be obtained from the analysis of each dataset in isolation.

5.6 False Discovery Rate

All participants of Dagstuhl Seminar 15351

Multiple testing has several contexts: Large number of statistical tests. What percentage of the rejected H0 are actually true? 'ome' assembly, i.e. assemble a shotgun sample, such as peptide and protein identification.

- **Statistical considerations.** Definition of FDR: expected proportion of false discoveries in the claimed set of discoveries. The keywords are 'discovery' (i.e., the definition of the experimental unit), and 'expected (i.e., this is an abstract concept that holds on average over an infinite replication of the experiment). Complications in proteomics: the experimental unit is not observed, but is inferred indirectly. The propagation of errors across the levels of integration (i.e. from spectra to peptides to proteins) has a lot of effect.
- **FDR estimation in microarrays.** Expect a mixture of uniform distribution and of a distribution around 0. Deviations from the uniform distribution can be due to violations of model assumptions within the experimental unit, or violation of independence between the experimental units.
- **FDR estimation in mass spectrometry.** In PSM identifications, the starting point is a score or a p-value. P-values are obtained by a generating function, separate decoy, concatenated target-decoy, or mix-max(?). Different null distributions may be needed for sequences of different uniqueness, some decoys look similar to true hits. Some applications require more stringent FDR cutoffs than others. An argument can be made for less stringent cutoffs in some cases.
- **Peptide and protein-level FDR.** Can be done by simulation, or by probabilistic modeling. A major problem is the fact that there are two different layers of uncertainty: in identification and in quantification. At the end biologists are interested in quantitative changes. How can we help them make decisions? They often do not appreciate the full extend of uncertainty. Most likely, the right decision will be made by considering various complementary, orthogonal types of experimental and prior information.

5.7 Correlation versus causality

Karen Sachs, Robert Ness, Kathryn Lilley, Lukas Käll, Sebastian Böcker, Naomi Altman, Patrick Pedrioli, Matthias Gstaiger, David Wishart, Lukas Reiter, Knut Reinert, Hannes Roest, Nicola Zamboni, Ruedi Aebersold, and Olga Vitek

Problem statement: Extract and mechanistically characterize the regulatory relationships in the biological system.

Biological challenges

- Regulatory relationships are large-scale and complex.

- Regulatory relationships are context-specific. The context can be spatial, temporal, or defined by interaction partners. A molecule (e.g. protein) can have different regulatory outcomes, depending on the context.
- Perturbations of a specific biochemical reaction or network (e.g., a protein KO) can have system-wide effects, beyond the target network.

Tools for inferring causal relationships. Regulatory networks are typically inferred from statistical associations between quantitative readouts. The networks are an intermediate step. Their goal is to suggest hypotheses for experimental follow up. The correct resolution (protein vs. protein complex vs. protein localization vs. protein PTMs . . .) should be chosen.

Statistical challenges. Statistical association can hide many types of causal events. Hidden aspects, which are not measured or not picked up by the model, complicate the task.

Big open question

- How to infer regulatory networks on a large scale?
- How to use networks to generate biological knowledge?

State of the art. Perturbations are key to elucidating causal events. Suppose we observe a statistical association between events A and B. To claim that A is a cause of B (i.e., A $\rightarrow$ B), we need to present a counterfactual argument that if A does not occur than B does not occur as well. This is best done by designing a perturbation experiment with and without A. The starting point is a statistical association. The association is often termed correlation. However, correlation strictly means linear association, and the reality is much more complex. E.g., if one protein deregulates another, the effect may not be a linear correlation, but a change in variability.

Statistical modeling. A statistical model of joint associations is needed, because humans cannot grasp the complexity, and can leap to erroneous conclusions too quickly. A combinatorial number of possible relationships is an issue. The required sample size (number of replicates) must grow super-exponentially to avoid spurious associations. The prior information (e.g., cell compartments, known functional associations) can impose constraints that can provide causality for the rest of the edges. All models are wrong, but some are useful. Correctness of a model is judged by how well it predicts the outcome of a new perturbation. The goal is to make the simplest model that explains the data.

Questions to address

- What is the available prior information?
- What is the minimal set of perturbations?
- How to incorporate the spatial and temporal context of the measurements? (Currently core models do not incorporate context).
- How can we understand the systems-wide effect of a perturbation, and extend the core models to the components beyond the target pathway? Since the effects of a perturbation are complex, small networks do not fully capture its effect, and prediction is ineffective.
- Effectively use of prior data (use weights / filter prior networks).

Suggestions to move forward. An iterative discovery process: start with seeking associations at a large scale to identify key players (and possibly reduce the list of components to be analyzed in detail), and follow up with targeted perturbation-based follow up experiments to look for causality among selected components The statistical formalism of the model can incorporate contextual annotations and constraints to scale the process, but the information

is not yet available, the sample sizes are small, and the computational complexity is large. Experts need to collaborate to put together the necessary components.

5.8 Metaproteomics

Josh Elias and Sven Nahnsen

Problem statement: Metaproteomes are immensely complex, and require new ways to process and evaluate data: standard proteomic strategies often do not scale.

Biological challenges

- Missing sample-specific metagenome: Unclear how to construct proteome database
- Dirty samples: Gel cleanup works, but is time-consuming, and may reject small, interesting
- Data integration: microbe enumeration, metagenome with proteomics
- Quantitation: how to normalize between heterogeneous samples? Searching "nr" database can be challenging: Search speed, FDR assessment at the protein AND organism level
- Sample storage conditions, like other body fluids, is a challenge for comparative studies
- Field collection also difficult to control
- Dietary components aren't readily identified with sequencing

Tools for metaproteome analysis

- MetaProteomeAnalyzer: Protein → Microbe mapping
- MetaProSIP: Analysis using stable isotope probing

Statistical /computational challenges

- peptide → protein → organism assignment (double FDR!!!)
- Distraction problem: When there's many more possible sequences than spectra available for matching, it's more likely for an incorrect match to out-rank a correct one

Big open question. What does metaproteomics get us that metagenomics does not?

Questions to address

- Health: What are potential antigens? How are microbes communicating with one another and with host (and how does this affect health)? Integration with disease biology: Make targeted assays? How do dietary proteins affect our intestinal immune surveillance?
- Systems Biology: Can we use the metaproteome to reduce the apparent complexity of the microbiota into more discrete functional (and manipulatable) modules? Many microbes make similar functional proteins or clusters of proteins; these functions may be more consistent between hosts than the microbes.
- Ecology: Non-gut communities are harder to assess: Oceans, soil, etc. Important aspects of ecosystems, but very poorly understood. (Mak Saito, WHOI)

State of the art

- Parallel metagenomic sequencing + proteomics; 6-frame translations (Banfield & Hettich)
- Large microbe databases + Organism assembly (MetaProteome Analyzer (Martens & Rapp)

Suggestions to move forward. Creation of reference datasets.

- In silico mixtures of discrete microbial proteomes (mono-culture datasets mixed post-acquisition).
- In vitro mixtures of known microbial cultures (mix microbial pellets at known, various concentrations).
- Co-culture of known microbes
- Dietary proteomes: more species to include in databases

5.9 Challenges in Quantitation

Jonathon O'Brien, Lukas Reiter, Susan Weintraub, Robert Chalkley, Rudolf Aebersold, Bernd Wollscheid, Pedro Navarro, Stephan Tate, Stefan Tenzer, Matthias Gsteiger, Patrick Pedrioli, Naomi Altman, and Hannes Röst

Statistical limitations/problems

- Peptide to protein rollup is a statistical inference problem
- There exists a wide variety of ad hoc methods → repeatability problems
- Different questions → different method
- Inconsistency of methods is an issue. On the other hand using the same methods for different technologies also creates problems
- Missing data is a problem. In Statistics missing data is generally categorized as missing completely at random (MCAR), missing at random (MAR) or non-ignorably missing. Non-ignorably missing data occurs frequently in proteomics experiments, meaning that the probability of being missing is directly dependent on the intensity value. This creates a bias.
- Pre-fractionation is difficult to handle. It doesn't have to be a problem but the variation in how software packages handle fractionation distorts the target of inference
- Ion suppression. Jonathon O'Brien mentions that he can see ion suppression. It was discussed whether there is really such a thing as ion suppression. If the samples are rather similar it is probably not a major issue. One can observe that the spray efficiency varies slightly over time but not dramatically.
- Misidentifications can cause both biases in point estimates and mis-labelled proteins.

Other limitations

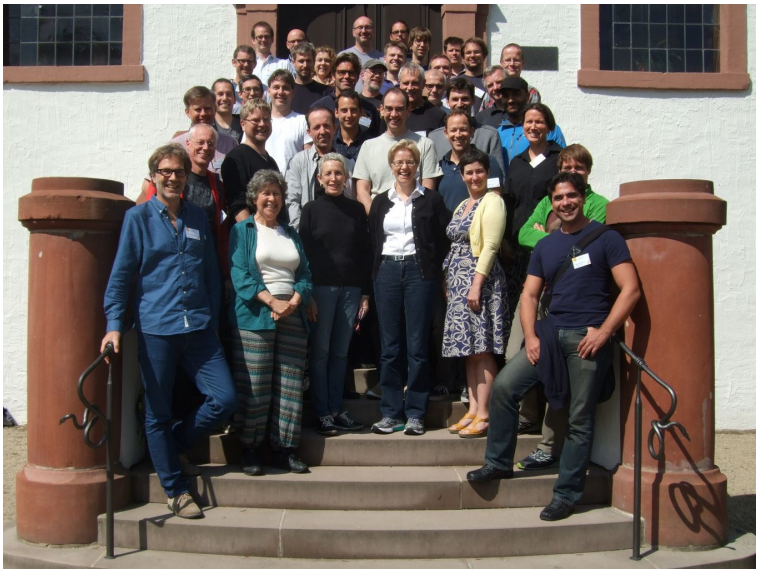
- Many samples and runs can be problematic → forces label free, which then puts further importance on normalization algorithms
- Quality control → quality of acquisition
- Making a statement on the protein quantity
- Housekeeping proteins. Naomi mentions that one houskeeper didn't work well for microarrays but using a panel of let's say 20 proteins worked quite well.
- Difference between nucleotide world is that the platforms are very homogenous → it's different in MS, there are distinct analyzers, different sample prep. methods
- Large experiments → make a note of the acquisition sequence to account for batch effects

Suggestions for progress

- Normalization in microarrays Affymetrix created a reference data set → everybody could try → eliminated a lot of methods from the field (it wasn't a formal process)
- It was suggested to make MS sessions at statistical conferences
- It was suggested to make a study comparing different quantitation strategies. Comparing different pipelines for the same workflow was already done and with encouraging results. Such studies have also been done in the microarray field
- ABRF was also a similar aim (only few instances of certain workflows)
- Methods that converted unrepeatable results to repeatable results are presented in Ting, L., Cowley, M.J., Hoon, S.L., Guilhaus, M., Raftery, M.J., and Cavicchioli, R. (2009). Normalization and Statistical Analysis of Quantitative Proteomics Data Generated by Metabolic Labeling. Mol Cell Proteomics 8, 2227–2242.
- Samples of e.g. three organisms mixed in different ratios can be used as benchmarking data sets
- Clinical tumor analysis consortium is setting standards. MSACL might be better suited to set standards. MSACL conference → clinical mass spectrometry → might be a good forum to present such a benchmarking study
- CPTAC study investigated how different labs can produce similar results when using their favourite method as compared a standard method. They only achieved consistent results with standardized workflows.

Participants

- Rudolf Aebersold
ETH Zürich, CH
- Theodore Alexandrov
EMBL Heidelberg, DE
- Naomi Altman
Pennsylvania State University – University Park, US
- Nuno Bandeira
University of California – San Diego, US
- Andreas Beyer
Universität Köln, DE
- Sebastian Böcker
Universität Jena, DE
- Robert Chalkley
University of California – San Francisco, US
- Joshua Elias
Stanford University, US
- Laurent Gatto
University of Cambridge, GB
- Anne-Claude Gingras
University of Toronto, CA
- Matthias Gstaiger
ETH Zürich, CH
- Kasper Daniel Hansen
Johns Hopkins University – Baltimore, US
- Henning Hermjakob
European Bioinformatics Institute – Cambridge, GB
- Michael Hoopmann
Institute for Systems Biology – Seattle, US
- Lukas Käll
KTH – Royal Institute of Technology, SE
- Oliver Kohlbacher
Universität Tübingen, DE
- Bernhard Küster
TU München, DE
- Kathryn Lilley
University of Cambridge, GB
- Lennart Martens
Ghent University, BE
- Sven Nahnsen
Universität Tübingen, DE
- Pedro José Navarro Alvarez
Universität Mainz, DE
- Robert Ness
Purdue University, US
- Jonathon O'Brien
University of North Carolina – Chapel Hill, US
- Patrick Pedrioli
ETH Zürich, CH
- Knut Reinert
FU Berlin, DE
- Lukas Reiter
Biognosys AG – Schlieren, CH
- Bernhard Renard
Robert Koch Institut – Berlin, DE
- Hannes Röst
Stanford University, US
- Karen Sachs
Stanford University, US
- Timo Sachsenberg
Universität Tübingen, DE
- Albert Sickmann
ISAS – Dortmund, DE
- Stephen Tate
SCIEX – Concord, CA
- Stefan Tenzer
Universität Mainz, DE
- Michael L. Tress
CNIO – Madrid, ES
- Olga Vitek
Northeastern University – Boston, US
- Christine Vogel
New York University, US
- Susan T. Weintraub
The University of Texas Health Science Center, US
- David Wishart
University of Alberta – Edmonton, CA
- Bernd Wollscheid
ETH Zürich, CH
- Nicola Zamboni
ETH Zürich, CH



Report from Dagstuhl Seminar 15352

Design of Microfluidic Biochips: Connecting Algorithms and Foundations of Chip Design to Biochemistry and the Life Sciences

Edited by

Krishnendu Chakrabarty[1], Tsung-Yi Ho[2], and Robert Wille[3]

1 **Duke University – Durham, US,** krish@ee.duke.edu
2 **National Tsing-Hua University – Hsinchu, TW,** tyho@cs.nthu.edu.tw
3 **University of Bremen/DFKI, DE, and Johannes Kepler University Linz, AT,** robert.wille@jku.at

Abstract

Advances in microfluidic technologies have led to the emergence of biochip devices for automating laboratory procedures in biochemistry and molecular biology. Corresponding systems are revolutionizing a diverse range of applications, e.g. air quality studies, point-of-care clinical diagnostics, drug discovery, and DNA sequencing – with an increasing market. However, this continued growth depends on advances in chip integration and design-automation tools. Thus, there is a need to deliver the same level of *Computer-Aided Design* (CAD) support to the biochip designer that the semiconductor industry now takes for granted. The goal of the seminar was to bring together experts in order to present and to develop new ideas and concepts for design automation algorithms and tools for microfluidic biochips. This report documents the program and the outcomes of this endeavor.

Seminar August 23–26, 2015 – http://www.dagstuhl.de/15352
1998 ACM Subject Classification B.7 Integrated Circuits,J.3 Life and Medical Sciences
Keywords and phrases cyber-physical integration, microfluidic biochip, computer aided design, hardware and software co-design, test, verification
Digital Object Identifier 10.4230/DagRep.5.8.34

1 Executive Summary

Krishnendu Chakrabarty
Tsung-Yi Ho
Robert Wille

Advances in microfluidic technologies have led to the emergence of biochip devices for automating laboratory procedures in biochemistry and molecular biology. These devices enable the precise control of nanoliter-scale biochemical samples and reagents. Therefore, *Integrated Circuit* (IC) technology can be used to transport a "chemical payload" in the form of micro- or nano-fluidic carriers such as droplets. As a result, non-traditional biomedical applications and markets (e.g., high-throughput DNA sequencing, portable and point-of-care clinical diagnostics, protein crystallization for drug discovery), and fundamentally new uses are opening up for ICs and systems. This represents a More than Moore-approach.

Design of Microfluidic Biochips: Connecting Algorithms and Foundations of Chip Design to Biochemistry and the Life Sciences, *Dagstuhl Reports*, Vol. 5, Issue 8, pp. 34–53
Editors: Krishnendu Chakrabarty, Tsung-Yi Ho, and Robert Wille

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Miniaturized and low-cost biochip systems are revolutionizing a diverse range of applications, e.g., air quality studies, point-of-care clinical diagnostics, drug discovery, and DNA sequencing. Frost & Sullivan recently predicted a 13.5% Compound Annual Growth Rate for the US biochip ("lab-on-chip") market during 2008-2015, and the market size for lab-on-chip alone (not including microarrays, biosensors, and microreactors) is expected to be over $1.6 billion in 2015. Similar growth is anticipated in other parts of the world, especially Europe and Japan. On a broader scale, the annual US market alone for in vitro diagnostics is as high as $10 billion and similar figures have been estimated for the drug discovery market. For clinical diagnostics, it has been predicted that we will soon see 15 billion diagnostic tests/year worldwide.

However, continued growth (and larger revenues resulting from technology adoption by pharmaceutical and healthcare companies) depends on advances in chip integration and design-automation tools. Thus, there is a need to deliver the same level of *Computer-Aided Design* (CAD) support to the biochip designer that the semiconductor industry now takes for granted. In particular, these CAD tools will adopt computational intelligence for the optimization of biochip designs. Also, the design of efficient CAD algorithms for implementing biochemistry protocols to ensure that biochips are as versatile as the macro-labs that they are intended to replace. This is therefore an opportune time for the software and semiconductor industry and circuit/system designers to make an impact in this emerging field.

Recent years have therefore seen growing interest in design methods and design-automation tools for the digital microfluidic platform, with special issues of *IEEE Transactions on CAD and IEEE Design & Test of Computers*, special sessions at *DAC*, *ISPD*, *ASPDAC*, and *ICCAD*, and workshops/tutorials at *ISCAS*, *ICCAD*, *SOCC*, and *DATE*. A number of CAD research groups worldwide (e.g., Duke University; Carnegie Mellon University; University of Texas at Austin; Rensselaer Polytechnic University; University of California at Riverside; University of Washington; Technical University of Denmark; Technische Universität München; University of Bremen; National Tsing Hua University; National Chiao Tung University, National Taiwan University; Tsinghua University; Indian Statistical Institute; Ritsumeikan University; Nanyang Technological University; Johannes Kepler University Linz) have initiated research projects on CAD for microfluidic biochips.

The goal of the seminar was to bring together experts in order to present and to develop new ideas and concepts for the design automation algorithms and tools for microfluidic biochips. Areas ranging from architecture, synthesis, optimization, verification, testing, and beyond have been covered. Topics which have been discussed included besides others:

- Architectural synthesis
- Behavior-level synthesis
- Cooling for integrated circuits
- Cross-contamination removal
- Cyberphysical integration
- Device modeling
- Drug-delivery biochips
- Fault modeling, testing, and protocol verification
- Light-actuated biochips
- Numerical simulation
- On-chip sensors
- Paper-based microfluidics
- Particle microfluidics
- Physical design

- Pin-constrained design
- Sample preparation

As results we received a better understanding of the respective areas, new impulses for further research directions, and ideas for areas that will heavily influence research in the domain of design automation on microfluidic biochips within the next years. The seminar facilitated greater interdisciplinary interactions between chip designers, bioengineers, biochemists, and theoretical computer scientists.

The high-quality presentations and lively discussions have been ensured by carefully selected experts who participated at the seminar. All of them have established for themselves a stellar reputation in the respective domains. While researchers working on design automation and optimization of microfluidic biochips build the majority of the participants, also some experts from surrounding research areas attended. For example, researchers working on emerging architectures and applications of microfluidic biochips provided the needed insight for the discussions about the practical problem formulation for commercialized product. Computer scientists with a focus on computer-aided design enriched the discussions about the top-down design methodology and optimization of large-scale components like mixers and routing channels. Therewith, the unique concept of Dagstuhl seminars was applied in order to bring researchers from different domains together so that the interdisciplinary topics could have been discussed and progress in these areas has been made.

2 Table of Contents

3 Overview of Talks

3.1 Hands-on Experiences on Actual Biochips

Mirela Alistar (Copenhagen)

Joint work of Alistar, Mirela; Trojok, Ruediger
URL http://www.bioflux.eu

It is not a secret that in biology laboratories hours of manual work are considered a compulsory part of the experiment. During a day of work, lab researchers have to pipette the right amounts of fluids in tubes, carry them from one machine to another, program and handle each machine individually, label and document carefully each step and then convert the results to data and analyze it. For a simple routine experiment, each of the mentioned tasks is performed at least 10 times. Past decade, a big effort has been done to produce machines (e.g., pipetting robots) that would automate some of the tasks in the lab. However, these machines were developed under the industrial mindset to maximize the throughput of a single task. Thus, these machines are of large size, task-specific, difficult to use (they usually come with dedicated drivers and software) and most importantly, extremely expensive.

Mirela Alistar and Ruediger Trojok are leading the BioFlux project, with the purpose to advance from automated biology to digital biology. In our vision, a digital lab should be: (1) fully integrated,running all the tasks on the same machine; (2) easy to use, with a web-based software for biological design of new experiments and hardware control; (3) general-purpose, allowing easy reconfiguration and design of new experiments; (4) cheap, offering open-source and do-it-yourself assembly kits.

During this workshop, we presented the common laboratory procedures for running synthetic biology applications. We showed a commercial DNA extraction kit (from Evogen Inc.) and presented the manual steps (pipetting, incubation, centrifuging) that the biologists have to take to extract the DNA.

Next, we emphasized that contamination is a significant issue by doing a microbiology experiment with one of the members of the seminary. We had talked about bacteria media, culture and growth. The participant was instructed to pour agar plates with LB-based media. After the plates set down, the participant went to wash his hands and then imprinted the plates with his fingers. The bacterial growth was monitored during the following days and all the seminar participants were updated on the progress.

The next part of the workshop consisted on a step-by-step instruction set on how to build your own biochip. We demonstrated using two versions of biochips: one that is manually controlled and one that is automatically controlled. The manual biochip was developed for a thorough study of the interaction between the fluids and the electrical potential. The automated biochip was controlled though an Arduino and can be programmed by any user due to the user-friendly Arduino interface. The seminary participants had all a chance to take photos, ask questions and test the biochips.

The last part of the workshop was dedicated to discussions. Some participants were interested in developing such low-cost DIY biochips in their groups for research purposes such as testing their own algorithms. Some discussions arose about funding possibilities, reliability of the DIY biochips, scalability of the products, end-users and applications.

Seminary participants from China, Korea, Taiwan and India expressed their desire to have such an workshop organized locally at their labs. Hence, the workshop resulted in follow-up discussions on grant applications.

3.2 Research on Biochip Architectures and EDA: Hype, Myths, and Reality (Reflections and Predictions)

Krishnendu Chakrabarty (Duke University – Durham, US)

Over the past fifteen years, significant research advances on design automation for microfluidic biochips have been reported. Early research was motivated by the considerable hype generated by technology demonstrations and the promise of a paradigm shift in molecular biology. While a sizeable research community has emerged worldwide and design automation for microfluidic biochips has now become an important component of major conferences (and the portfolio of the top journals) in the area, skepticism continues to be voiced about the practicality of design automation solutions and the relevance of this research to the broader community of biochip users. In this talk, I presented a retrospection of the early hype and some of the myths that have been exposed. A snap poll of the audience was taken with respect to a series of controversial questions. I also highlighted specific problems that design automation must tackle and led a discussion on how our community can engage in a more meaningful way with life science researchers. The discussion was lively and highly interactive. At the end, we collectively identified strategies for advancing from manipulating small volumes of liquid on a chip to accomplishing realistic biochemistry on these chips.

3.3 On-chip Logic Using Pneumatic Valves

William H. Grover (University of California at Riverside, US)

Microfluidic chips are capable of performing a wide variety of different applications faster, cheaper, and better than conventional lab-scale tools. However, the spread of microfluidic technologies is slowed by the amount of off-chip hardware required to operate microfluidic chips. This off-chip hardware is often far more expensive, bulky, and power-hungry than the chip itself, a fact that makes microfluidic instruments less suitable for use in resource-limited or point-of-care contexts. Here I describe how off-chip hardware can be reduced or eliminated by integrating the control of a microfluidic device onto the chip itself. We accomplish this using *monolithic membrane valves*, pneumatically-actuated microfluidic valves that we originally developed for controlling fluid in microfluidic chips. After finding that these valves can control air flow as well (and thereby control each other), we developed an assortment of valve-based logic gates and circuits. These principles of pneumatic logic are powerful enough to control even the most complex microfluidic chips using little or no off-chip hardware. Designing these valve-based logic circuits is not trivial, but automating their design could be a fertile area of inquiry for researchers working on microfluidic design automation.

3.4 Integrated Fluidic-Chip Co-Design Methodology for Digital Microfluidic Biochips

Tsung-Yi Ho (National Tsing-Hua University – Hsinchu, TW)

Recently, digital microfluidic biochips (DMFBs) have revolutionized many biochemical laboratory procedures and received much attention due to many advantages such as high throughput, automatic control, and low cost. To meet the challenges of increasing design complexity, computer-aided-design (CAD) tools have been involved to build DMFBs efficiently. Current CAD tools generally conduct a two-stage based design flow of fluidic-level synthesis followed by chip-level design to optimize fluidic behaviors and chip architecture separately. Nevertheless, existing fluidic-chip design gap will become even wider with a rapid escalation in the number of assay operations incorporated into a single DMFB. As more and more large-scale assay protocols are delivered in current emerging marketplace, this problem may potentially restrict the effectiveness and feasibility of the entire DMFB realization and thus needs to be solved quickly. In this research, we propose the first fluidic-chip co-design methodology for DMFBs to effectively bridge the fluidic-chip design gap. Our work provides a comprehensive integration throughout fluidic-operation scheduling, chip layout generation, control pin assignment, and wiring solution to achieve higher design performance and feasibility. Experimental results show the effectiveness, robustness, and scalability of our co-design methodology on a set of real-life assay applications.

3.5 Sample Preparation on Microfluidic Biochips

Juinn-Dar Huang (National Chiao-Tung University – Hsinchu, TW)

My recent research direction is about sample preparation in microfluidic biochips. Sample preparation on microfluidic chips actually refers to a set of problems, which can be classified in different perspectives. For example, the optimization goal can be reactant minimization, operation count minimization, waste minimization, and so on. The target microfluidic biochip can be digital (1-to-1 mixing model only) or flow-based (a mixer with N segments, $N > 2$) as well. The target concentration value of product solution can be just single one or multiple at the same time. The number of reactants in a bioassay can be at least two or more. Each different combination of the aforementioned parameters defines a unique sample preparation problem and needs to be properly solved. In the meantime, I am currently working on so-called cyber-physical sample preparation technology, which can dynamically adjust the preparation process based on real-time on-line feedback.

3.6 Using Boolean Satisifability to Design Digital Microfluidic Biochips

Oliver Keszöcze (DFKI – Bremen, DE)

Joint work of Keszöcze, Oliver; Wille, Robert; Chakrabarty, Krishnendu; Drechsler, Rolf

Advances in microfluidic technologies have led to the emergence of Digital Microfluidic Biochips (DMFBs), which are capable of automating laboratory procedures in biochemistry and molecular biology. During the design and use of these devices, droplet routing represents a particularly critical challenge. Here, various design tasks have to be addressed for which, depending on the corresponding scenario, different solutions are available. However, all these developments eventually resulted in a huge variety of different design approaches for routing of DMFBs – many of them addressing a very dedicated routing task only.

In this presentation, we show a comprehensive routing methodology which

1. provides one (generic) solution capable of addressing a variety of different design tasks,
2. employs a "push-button"-scheme that requires no (manual) composition of partial results, and
3. guarantees minimality e.g., with respect to the number of timesteps or the number of required control pins.

The approach is not to find an algorithm that solves every possible routing problem but to formally model biochips and corresponding routing problems and give that to a SMT solver (Z3 in our case) which then, in turn, produces a routing solution. This formal model consists of diverse variables that describe the system's states and constraints on these variables that model how the droplets may move as well as constraints such as the fluidic constraints.

One exemplary constraint is for the actual movement of droplets. Our approach models the movements in a backward manner. The constraints for the presence of a droplet in a specific positions means that the droplet must have been present in the neighborhood of that position in the previous time step.

This routing process then is done in an iterative manner:

1. set T to 0
2. create the model that spans T time steps
3. ask the solver to find a routing solution for that model
4. if no solution is found, increase T by one and go to 2)

Finding a solution in such a manner has two desired properties:

1. the solution is guaranteed to be minimal with respect to the amount of ime steps used in routing
2. the solution is definitely valid in the model.

In the presentation we show how to easily extend the model to consider many different aspects (e.g. fluidic constraints, pin assignment). The good thing of our approach is that there is no need to think of how the newly added problem is to be solved (the only thing to be done is to add a parameter for the amount of pins P in the iterative process described above). The solver does the main work in the background. This works especially well when to separate but interconnected tasks (i.e. droplet routing and pin assignment) are solved at the same time; no propagation of information between two different problems has to be performed by the developer.

3.7 Demo of a Visualization Tool for Digital Microfluidic Biochips

Oliver Keszöcze (DFKI – Bremen, DE)

Joint work of Keszöcze, Oliver; Stoppe, Jannis; Wille, Robert; Drechsler Rolf

There are various challenges in the development of digital microfluidic biochips Design tasks such as synthesis, routing and layouting are complex and currently being investigated by various research institutes in their ongoing endeavours. However, so far there is no tool to easily visualize the results of given approaches, making the development and analysis of approaches for these chips a tedious task.

We present a visualization engine to display a given microfluidic biochip design (e.g. the routings paths for given nets). The visualization is supposed to be easy to use, resulting in a hassle-free environment for designers to work in.

Additionally to displaying static information such as grid layout droplet and dispenser/sink position we support to visualize dynamic information such as droplet and mixer positions as well as cell actuations. To help the developer in the analysis process, the transitions between time steps (i.e. system states) is animated. This greatly helps to understand where a certain droplet game from at any given time step; this is especially helpful when moving many droplets at once. Further more, the tool displays the aggregated information of the droplet positions (i.e. the paths droplets take).

The tool has been implemented in Java using the libgdx library. Java eases the process of deploying the tool as it should run out of the box on all mayor system supporting Java. The libgdx library uses the full power of OpenGL, allowing to

(a) easily animate the system and

(b) smoothly zoom and scroll through the system under inspection.

3.8 Biochips: The Wandering of an EDA Mind (A Case Study)

Bing Li (TU München, DE)

Joint work of Li, Bing; Schlichtmann, Ulf; Ho, Tsung-Yi
Main reference T.-M. Tseng, B. Li, T.-Y. Ho, U. Schlichtmann, "Reliability-aware Synthesis for Flow-based Microfluidic Biochips by Dynamic-device Mapping," in Proc. of the 52nd Annual Design Automation Conf. (DAC'15), Article No. 141, 6 pages, ACM, 2015.
URL http://dx.doi.org/10.1145/2744769.2744899

Microfluidic biochips have revolutionized traditional biochemical diagnoses and experiments significantly by exactly manipulating nanoliter samples and reagents. This miniaturization saves expensive reagents and improves experiment accuracy effectively. With the recent advances in manufacturing technology and integration of biochips, very complex applications can now be executed on such a chip as a whole without human interference. However, the interface between this tremendous engineering advance and applications is still incomplete, and new emerging architectures are enlarging this gap further.

In this presentation, we discuss the challenges in mapping applications to several newly emerged biochip architectures. We first explain a method to improve the reliability of flow-based biochips by assigning devices dynamically on a fully programmable valve array. In a traditional flow-based biochip, valves that drive or pump fluid samples in mixers actuate

10 times more than the other valves that control flow transportation. Since the entire chip fails when any of these valves wears out, this imbalance of actuations affects the lifetime of the chip significantly. To alleviate this problem, we allow valves to change their roles during the execution of an application. In this concept, valves that have been used to pump fluid samples in a mixing operation are used to control flow transportation thereafter. Because the valves along a mixer have different roles in different mixing operations, valve actuations are distributed more evenly. Consequently, the maximum number of valve actuations in executing an application can be reduced effectively without incurring any additional cost.

In addition to reconfigurable valve arrays, we also discuss biochips printed on paper and biochips with capacitors that control electrodes in a row-column refreshing mode. Challenges in adopting paper-based biochips come from the fact that electrodes are printed only on one side of the paper. Therefore, wires providing voltages to electrodes must be routed at the same layer. The other new biochip architecture with control capacitors is based on the thin film transistor (TFT) technology. In such a chip, each pixel on the LCD plane has a capacitor-like cell. To set voltages to some pixels, a row-column write process sweeps all the capacitance cells. To use TFT pixels to manipulate droplets, challenges still remain. The first one is actually from the extremely refined electrodes. To move large droplets, multiple electrodes should be grouped dynamically for operations and transportation. The second challenge is that the voltages to electrodes should be set in the row-column mode instead of independently. It should be guaranteed that the voltage setting process does not affect the droplets on the chip.

By exploring several emerging biochip architectures, we have demonstrated challenges in mapping applications to them. To achieve a wide adoption of these new architectures in industry, a close collaboration between the chip design community and the EDA community is indispensable.

3.9 Microfluidic Large-Scale Integration and its Applications in Life Science

Sebastian J. Maerkl (Ecole Polytechnique Federale de Lausanne (EPFL), CH)

The Dagstuhl seminar #15352 on "Design of Microfluidic Biochips: Connecting Algorithms and Foundations of Chip Design to Biochemistry and the Life Science" brought together a group of scientists with backgrounds in computer science, microengineering / microfluidics, and researchers familiar with biochemistry. The Dagstuhl conference began with seminars to provide background and up to date information on the state of the various research fields represented at the conference. The second day was dominated by short as well as in depth tutorial and discussion sessions. Informal discussions were possible throughout the duration of the meeting.

I contributed an approx. 45 minute seminar providing a short review of multilayer soft lithography [1] and microfluidic large-scale integration [2], as the conference primarily focused on electrowetting based digitial microfluidic devices. This short technical introduction on MLSI was followed by a description of a proof-of-concept programmable valve-based microfluidic device including an explanation of the basic design concept, the technical developments required to achieve sufficiently high chip complexity, and the implementation of

basic fluidic operations such as on-the fly device reprogramming, fluid metering, and mixing [3]. This description was followed by two basic, proof-of-concept biological applications. The first application described the implementation of a standard immunoassay on the platform, which is commonly employed in a plethora of clinical diagnostic assays. The second application showed that the reconfigurable device could be applied to cell manipulations and on-chip culturing using *S. cerevisiae* as the model system.

During the second half of the seminar I described our recent efforts at developing methods and tools for cell-free synthetic biology. We recently developed a microfluidic chemostat device with a parallel architecture [4]. This devices allowed us to run *in vitro* transcription / translation (ITT) reactions at steady-state for up to 30 hours. Previously, such reactions were run in standard batch reaction format in test tubes, which severely limited the usability of ITT reactions for the implementation and characterization of genetic networks. Previously, only genetic cascades had been implemented in ITT reactions, primarily due to these technical restrictions [5]. With our novel microfluidic chemostat arrays we could show that genetic oscillators could be successfully implemented in a cell-free environment. We also showed that a diverse set of native biological regulatory mechanisms could be reconstituted on this platform. We then went on to show that the platform could successfully implement the repressilator, a classic synthetic network, which had been designed and implemented in *E. coli* [6]. We went on to show that we could rapidly characterize biological parts and devices, creating novel 3-node as well as the first 5-node genetic oscillators. To demonstrate that novel genetic networks implemented and optimized *in vitro* could be transferred to a cellular environment, we transferred both our 3-node and 5-node genetic oscillators to *E. coli* and characterized these two networks on the single cell level; proving that transfer is possible. We also discovered that our 3-node genetic oscillators were surprisingly synchronous as opposed to the original repressilator networks. Our current working hypothesis explaining the difference in phenotype of these two genetic oscillators, which share the exact same network architecture, lies in the fact that the molecular concentrations in our newly engineered genetic networks is likely higher than those of the original repressilator, leading to a drastic reduction in noise, which in turn results in synchronized behavior of cells in the same lineage.

Current Problems in Microengineering / Microfluidic Device Design and Implementation

Dagstuhl represented an opportunity to meet the biochip EDA community, with whose work I was only marginally familiar. This fact is probably telling and represents a significant gap between the community of microengineers / chip developers and the community of EDA computer scientist who are working on problems related to biochip design and biochip operation. The fundamental problem facing the biochip community at the moment lies in bridging the gap between developing and working on "toy" problems, which remain of low interest to the microengineering community. A similar challenge exists for the microengineering community who build new microfluidic tools and biologists and chemists for whom these new tools are being developed. It may thus be insightful to briefly describe the current challenges facing the microfluidic / microengineering community, and to describe possible approaches to maximize the impact microengineers can achieve through their work.

The first challenge for microengineers developing new microfluidic devices is to identify well-known shortcomings or limitations of currently existing technologies and to develop a novel approach that leads to a significant improvement in performance relevant for the end-users of the technology. An alternative approach involves identifying an area of biology in which no methods are available, but a clear and obvious need for novel methods exists. Development of microfluidic single cell approaches represents this second approach. A clear

set of biological questions existed, but no technologies exited that could be employed to answer these questions, or the existing methods were insufficient. A third possible strategy to develop methods for biology is to enable an entirely novel approach to conducting biology which does not serve a pre-existing community of biologists, but around which a group of scientists will form because it provides a unique and novel way of approaching a problem in biology (synthetic biology, and now cell-free synthetic biology could be considered such fields). Selecting an appropriate problem to solve unfortunately only represents the first step in the process. Even if one develops the methods and tools, impact will remain limited to the community of microengineers, and will fail to impact the biological community unless the second challenge is addressed as well.

The second challenge represents the difficulty of impacting the intended end-users of the technology (in this case biologists). This problem exists because it has proven extremely difficult to transfer microfluidic technology to biology laboratories. The entrance barrier to the field of microfluidics is sufficiently high to prevent a majority of biological labs from adopting this technology. Some biological communities have made more significant efforts in adopting microfluidic technologies and are actively involved in their development. The microbiology community is probably the community, that has made the most significant efforts in this, probably because it is obvious to this community that a number of central questions in the field of microbiology will only be answerable if microfluidic devices are employed. Most other biological communities have adopted microfluidic technology only if a commercially available system exits that meets their needs. For example, the CellAsic platform is fairly popular with microbiologists, and the Fluidigm single cell analysis devices developed for mammalian cells fill a clear need for current cell biological research. Other fields in which microfluidics is likely going to have significant impact is in personalized medicine and personalized diagnostics, through the development of next generation diagnostics platforms.

There are thus two possible approaches that can be taken if a microfluidic technology is to significantly impact the biological community. The first approach requires that technologies developed in the lab are ultimately commercialized either through start-up companies or through licensing to existing companies. In many instances, significant additional development is required to make novel microfluidic tools and methods sufficiently user friendly to allow commercialization. This area unfortunately represents a difficulty in that it rarely is of interest to an academic lab to conduct such engineering work, and in many instances investors seem to prefer more mature technologies for funding. Technology transfer is thus necessary and of utmost importance in order to maximize impact of microfluidic technologies, but is also extremely difficult.

An alternative to transferring novel technology to the commercial sector so that it becomes accessible and usable by the biology community is to directly conduct biologically relevant experiments with the newly developed technology. Impact in biology can be achieved by supplying biological datasets or novel biological insights in the form of new mechanisms, the discovery of novel molecules, or discovering novel links between existing molecules, which represent the goals classically pursued by biologists. Providing quantitative information on otherwise well known or well characterized molecules can also have considerably impact in biology [7, 8], as precise and comprehensive data can challenge existing biological dogmas derived previously based on low-quality experimental data, limited by technologies available at the time. Finally the development and characterization of synthetic biological systems on all levels is of high interest, and novel technologies are expected to facilitate such developments. The unifying characteristic of these foci is that data or biologically molecules can be easily shared with biological laboratories, and can thus readily impact biological research.

Challenges and Opportunities for Computer Science in Microengineering and Biology

The reason for describing the challenges facing the microengineering / microfluidics community is that the computer science / EDA community currently working on microfluidic devices is facing similar challenges. In order for the EDA community to impact the microengineering / microfluidics community, or the biology community requires that relevant problems are being identified and solved, and that these solutions are immediately accessible and usable by the target communities of researchers. As microfluidic platforms are becoming commercially available, biologists will likely adopt them if they provide a performance advantage over existing approaches. These advantages could be any combination of decreased cost, increased throughput, and automation. In addition, applications of microfluidic devices in the near future will remain task specific. In other words, biologists will conduct a particular task, or workflow, on a microfluidic platform such as molecular cloning, single cell analysis, or biochemical analysis. Furthermore, these tasks will generally follow a fairly well defined protocol and series of steps, with the only difference between experiments being the reagents/cells used on the devices. These requirements can be either fulfilled by valve-based or electrowetting microfluidic devices, but does not necessarily require a completely reprogrammable microfluidic device. The complexity of the needed control software therefore is likely to remain fairly limited in the foreseeable future.

Current opportunities for EDA based design and related approaches derived from computer science at the interface of microengineering / biology include the development of user-friendly control interfaces for electrowetting devices. As these devices become commercially available, better control software is needed that allows biologists to easily program their own routines on these chips. Such control software could provide an easy to use interface to defined fluid handling on the devices, and/or can be supported by more sophisticated protocol optimization algorithms. Similar control problems likely also exists for large, central robotic facilities, in which optimization is a non-trivial task. It might be of interest to contact big pharma companies to assess their needs in this domain. Finally, synthetic biology is currently facing considerable difficulties in developing rational approaches for biological network design. Although these networks still remain fairly simple, even simple networks require computer modeling to assess and optimize their performance. One the one hand, this situation is expected to become much more difficult as network size continues to grow. But, at the same time the underlying computational models and parts characterization will also drastically improve, allowing more accurate predictions to be made. There is also a clear precedence for the need and usefulness of extremely complex and sophisticated networks as found in any naturally occurring organism. It is thus almost inevitable that all biological network design in the present as well as in the near future, should or will be conducted *in silico*.

References

1 Unger, M. A., Chou, H. P., Thorsen, T., Scherer, A. & Quake, S. R. Monolithic microfabricated valves and pumps by multilayer soft lithography. Science 288, 113–116 (2000).
2 Thorsen, T., Maerkl, S. J. & Quake, S. R. Microfluidic large-scale integration. Science 298, 580–584 (2002).
3 Fidalgo, L. M. & Maerkl, S. J. A software-programmable microfluidic device for automated biology. Lab Chip 11, 1612–1619 (2011).
4 Niederholtmeyer, H., Stepanova, V. & Maerkl, S. J. Implementation of cell-free biological networks at steady state. Proc Natl Acad Sci USA 110, 15985–15990 (2013).
5 Noireaux, V., Bar-Ziv, R. & Libchaber, A. Principles of cell-free genetic circuit assembly. Proc Natl Acad Sci USA 100, 12672–12677 (2003).

6 Niederholtmeyer, H. et al. A cell-free framework for biological systems engineering. bioRxiv (2015). doi:10.1101/018317
7 Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. Science 315, 233–237 (2007).
8 Rajkumar, A. S., Denervaud, N. & Maerkl, S. J. Mapping the fine structure of a eukaryotic promoter input-output function. Nat Genet 45, 1207–1215 (2013).

3.10 Programming and Physical Design Tools for Flow-based Biochips

Paul Pop (Technical University of Denmark – Lyngby, DK)

Joint work of Pop, Paul; Madsen, Jan; Hassan Minhass, Wajid
URL http://www2.compute.dtu.dk/~paupo/talks/paupo_dagsthul15.v1.pdf

Microfluidic biochips are replacing the conventional biochemical analyzers by integrating all the necessary functions for biochemical analysis using microfluidics. Biochips are used in many application areas, such as, in vitro diagnostics, drug discovery, biotech and ecology. The focus of this special session is on continuous-flow biochips, where the basic building block is a microvalve. By combining these micro valves, more complex units such as mixers, switches, multiplexers can be built, hence the name of the technology, "microfluidic Very Large Scale Integration" (mVLSI). This talk has presented methods and tools for the programming and physical design of mVLSI biochips.

3.11 Algorithms for Automated Sample Preparation using Digital Microfluidic Biochips

Sudip Roy (Indian Institute of Technology – Roorkee, IN)

Main reference S. Roy, B. B. Bhattacharya, S. Ghoshal, K. Chakrabarty, "Theory and analysis of generalized mixing and dilution of biochemical fluids using digital microfluidic biochips," ACM Journal of Emerging Technologies in Computing Systems, Vol. 11, Issue 1, Article No. 2, 33 pages, ACM, 2014.
URL http://dx.doi.org/10.1145/2629578

In the last two decades, an emerging technology of "Lab-on-a-Chips (LOCs)" has been studied by the researchers of interdisciplinary fields to develop microfluidic biochips that can implement wide-range of biochemical laboratory test protocols (a.k.a. bioassays). A marriage of microelectronics and in-vitro diagnostics areas has led to this field of interdisciplinary research around LOCs or microfluidic biochips. In contrast to continuous-flow microfluidic chips, digital microfluidic (DMF) biochips are of a popular kind of microfluidic LOCs that can implement bioassays on an electrode array of a few square centimeters in size by manipulating micro/nano/pico liter volume fluid droplets. The functionality of a DMF biochip includes the following operations: dispensing the desired amount of fluids to the chip from the outside world as droplets, transporting the droplets on-chip to appropriate locations, mixing and splitting of several droplets, executing a well-defined bioassay on-chip, and finally analyzing the results at an on-chip detection site. Recent years have seen a surge in interest in design automation methods for DMF biochips. Along with several synthesis steps of DMF biochips

(like Scheduling, Module Binding, Placement, Droplet Routing, Wire Routing), protocol derivation for automatic sample preparation (dilution & mixing) using DMF biochips.

Our research envisions the algorithmic microfluidics and it expands the computer-aided-design (CAD) research to develop DMF biochips by designing algorithms for automated sample preparation (dilution and mixing) on such chips. Mixing and dilution of fluids are fundamental preprocessing steps in almost all biochemical laboratory protocols. Mixing of two or more fluids with a given ratio is often required as a preprocessing step of many real-life biochemical protocols, e.g., polymerase chain reaction (PCR). Dilution of a biochemical fluid is the special case of mixing, where only two different types of fluids, one of which is a buffer solution, are mixed at a certain ratio corresponding to the desired concentration. The dilution is commonly used in biological studies to create a variety of concentrations of the stock solution by mixing it with its diluents and it is required for sample preparation in many bioassays, e.g., real-time PCR, immunoassays, etc. For high-throughput applications, it is a challenge to determine the sequence of minimum number of mix-split steps for on-chip sample preparation. Furthermore, the production of waste droplets and/or the reactant fluid usage should be minimized. Moreover, design automation tools are necessary for optimizing the layout of the biochips.

In Dagstuhl seminar, we discussed about the basic background of DMF biochips and about several algorithms and CAD techniques for automated and on-chip fluidic sample preparation (dilution and mixing) of biochemical fluids using DMF biochips. We expect that for the betterment of our society, several low-cost, portable, automated biochemical laboratory-on-a- chips will be developed soon. In order to conduct innovative and basic research in developing of DMF biochips, it requires joining hands of experts from multiple disciplines: Computer Science, Electronics, Mechanical, Chemical, Biomedical Engineering, Microfluidics Sensor Technologies, Medical Science, etc.

3.12 Active Digital Microfluidic Paper Chips with Inkjet-printed Patterned Electrodes and their Point-of Care Biomedical Application

Kwanwoo Shin (Sogang University, KO)

Recently, our group has presented a novel paper-based fluidic chip that can enable the full range of fluidic operations by implementing an electric input on paper via an electrowetting technique [1, 2]. This powered paper-based microfluidic chip, which is known as an active paper open chip (APOC), is primarily characterized by discrete drop volumes and is an open-type chip. These active, paper-based, microfluidic chips driven by electrowetting are fabricated using inkjet printing technique and demonstrated for discrete reagent transport and mixing [1]. Instead of using the passive capillary force on the pulp in the paper to actuate a continuous flow of a liquid sample, a single, discrete drop or a group of digital liquid drops are perfectly transported along programmed trajectories. The patterned electrodes, which are designed on a desktop computer, are printed on low-cost paper, such as recycled magazine papers, with conductive CNT ink using an office inkjet printer [2], which should enable true point-of-care production and diagnostic activities. I presented our newly developed active paper open chips and their biomedical application. The solution simplifies the workflow and improves the reaction accuracy tremendously.

References

1 Hyojin Ko, Jumi Lee, Yongjun Kim, Byeongno Lee, Chan-Hee Jung, Jae-Hak Choi, Oh-Sun Kwon, Kwanwoo Shin, Active Digital Microfluidic Paper Chips with Inkjet-Printed Patterned Electrodes, Advanced Materials, 26, 2335–2340 (2014)

2 Oh-Sun Kwon, Hansu Kim, Hyojin Ko, Jumi Lee, Byeongno Lee, Chan-Hee Jung, Jae-Hak Choi, and Kwanwoo Shin, Fabrication and characterization of inkjet-printed carbon nanotube electrode patterns on paper, Carbon (2013) 58, 116–127

3.13 Bioflux Technology Dagstuhl Report

Rüdiger Trojok (KIT – Karlsruher Institut für Technologie, DE)

Digital Biology

Recent advances in Synthetic biology open up new possibilities in healthcare, agriculture, chemicals, materials, energy, and bioremediation. To date this is still a very labor intensive task that requires skilled technicians and scientists. However, manual work is time consuming and wages drive development costs, thereby restricting possibilities for rapid prototyping in synthetic biology. Digital Biology is the computer aided programming of biological assays using digital microfluidic biochip devices based on electrowetting on dielectric technology. Advanced laboratory hardware will make access to biotechnogical procedures much more affordable with easy to replicate 'Do It Yourself' equipment, further also increase automation, replace time consuming labour and increase replicability and standardisation of methods. Thus, Digital Biology allows for wide scale automation of laboratory procedures in synthetic biology by improving efficiency between 1000 to 100000 fold compared to manual laboratory work, for the first time enabling wide scale rapid prototyping for the iterative creation of biological systems. This will allow even small biological laboratories in academia and industry as well as researchers in the developing world to develop synthetic biology products.

Bioflux Technology

To successfully decentralize the Digital Biology technology, we want to develop Bioflux Technology—a platform that will automate the synthetic biology flow with great medical and commercial potential. Bioflux Technology will be a combination of a software suite for biologists to plan experiments. Microfluidic device, electronics hardware to run the experiments and the required wetware (biological reagents) to perform a wide range of standardized bioassays used in synthetic biology. The hardware consists of computer controlled microchips which switch on high DC voltage on a set of electrodes. The electrodes will be printable on superhydrophically coated paper. The layout of the papers is customized to the specific bioassay. The papers can be exchanged, while the hardware setup remains the same thus avoiding contamination issues in the bioassays.Only the program in the computer and the wetware on the paper is actualized for every use case. The main users of the technology are thought to be medical personnel and biologists for field diagnostic and health treatment applications. As soon as we have developed a device that is robust and compact, the use of Bioflux Technology can be extended to a large mass of users, such as farmers (for plant treatment) or regular citizens (for rapid point of care testing). Users will be able to use Biofux Technology to design and test their desired protocols, at low cost (provided by the

small scale of biological material used) and at faster speed (enhanced by microfluidics). Bioflux Technology will be cheap, easy to distribute around the world, usable on-site where the samples are taken and connected to a global database for further analysis of the sampled data.

To render Digital Biology accessible for synthetic biology, all fundamental biological assays used in synthetic biology need to be downscaled in volume and properties to function on a Bioflux platform. This entails protocols for in vitro DNA replication and assembly, protein expression and purification and cell transformation and incubation. The fundamental assays will be integrated into composite protocols applied in synthetic biology, depending on customer needs. Each protocol will be adapted for execution on the Bioflux platform and made controllable by our specially designed software. Protocols can be flexibly created out of fundamental assays in an online user interface with a customer protocol designer. The protocols will then be loaded onto the Operating System of a Bioflux platform. The user of the device then needs to load the for the protocol required wetware input on the chip. After activation, the operating system will execute the protocol and put out the desired wetware to a designed position on the chip. Wetware output could be synthetic assembled genomes, designer proteins, cells or secondary metabolites such as specialty chemicals. Besides, the software can output measurement signals of the conducted reactions, allowing for use in medical diagnostics. The Bioflux team favours open source innovation and a global collaboration with academic and non academic partners to advance the field of Digital Biology together.

Use case: Biostrike

An overuse of the available antibiotics and subsequent evolutionary pressure led to the development of multi-resistant bacteria. By now, the situation is becoming urgent, as very few effective drugs are left to treat infections. Antibiotic resistance development is a natural process. Bacteria are under selective pressure and evolve mechanisms to avoid the antimicrobial effects of the antibiotics. Once developed, the genes for the resistance then rapidly spread even cross over between different species – a process called horizontal gene transfer. It therefore is necessary to continuously develop new antibiotics to keep up pace with resistant bacteria. However, in 1990 there were 18 companies developing new antibiotics, by 2011 there were only 4. In 1990 10 new antibiotics were licensed, in 2011 only 2. The reason for a worsening of the antibiotics problem into an antibiotics crisis is a classical market failure because there is a lack of financial incentives for the pharmaceutical industry to involve in the development of drugs like antibiotics with a small profit margin. Decentralizing the screening for antibiotics around the world using cheap and fast Digital Biology could provide a solution to this problem. On one hand to reduce the costs of research allowing more people could contribute to find a common solution and on the other hand to increase the chances to discover new compounds. In a citizen science project, people around the globe could contribute to the solution of the antibiotics problem by identifying new antibiotics in a crowd-sourced research approach using Bioflux Technology. Specialists from all fields of expertise could design the bioassays for Point of care diagnostic and treatment of multiresistant bacteria. In practise, resistant bacteria could be collected by medical personnel, screened with the Bioflux platform and the results gathered in a central online database. The databases would be accessible to a global community of researchers that shares the task to design a case specific treatment. By rational and creative design of for example Bacteriophages, entirely new antibiotics could be designed. Bacteriophages are programmable macromulecules that specifically target a multiresistant bacteria strain. To date, they can be

readily designed using synthetic biology methods. Ultimately, only the clinical trials would have to be organized by a central agency, while all other steps of the diagnosis, finding the right cure and even the production of the antibiotics could be done in a decentralized and global collaboration of scientists.

3.14 Scalable One-Pass Synthesis for Digital Microfluidic Biochips

Robert Wille (University of Bremen/DFKI, DE, and Johannes Kepler University Linz, AT)

Digital Microfluidic Biochips (DMFBs) have been proposed to automate laboratory procedures in biochemistry and molecular biology. The design of the corresponding chips received significant attention in the recent past and is usually conducted through several individual steps such as scheduling, binding, placement, and routing. This established scheme, however, may lead to infeasible or unnecessarily costly designs. As an alternative, one-pass-synthesis has recently been proposed in which the desired functionality is realized in a single design step. While the general direction is promising, no scalable design solution employing this scheme exists thus far. In this work, we address this gap by proposing an automatic design approach which follows the one-pass synthesis scheme, but, at the same time, remains scalable and, hence, applicable for larger designs. Experiments demonstrate the benefits of the solution.

3.15 Flow-based Microfluidic Biochips

Hailong Yao (Tsinghua University – Beijing, CH)

Microfluidic biochips have emerged to revolutionize the traditional biological, biochemical and biomedical experimental processes. Noticeable merits of microfluidic biochips over traditional laboratory platforms include: (1) greatly saving the assay cost by reducing expensive samples/reagents to nano-liter or pico-liter volume, (2) effectively integrating the automatic control logic for reduced human intervention and labor cost, (3) significantly increasing sensitivity, accuracy and throughput, (4) essentially facilitating portability for point-of-care diagnostics, and (5) naturally enabling microscale assays (e.g., single-cell culture, capture and analysis) that are infeasible by traditional macroscale approaches. According to Research and Markets, the global biochips market is expected to grow at a CAGR of 18.6% from 2012 to 2018, and will reach $11.4 Billion by 2018. Applications of biochips cover many different fields, such as diagnostics and treatment, drug discovery and development, biological research, forensic analysis, agriculture, environmental sensors, food inspection, etc.

Flow-based microfluidic biochips are among the most commonly used microfluidic biochips both in laboratories and hospitals. Flow-based microfluidic biochips typically consist of several functional layers, which are fabricated by elastomer material (polydimethylsiloxane, PDMS) using the multilayer soft lithography (MSL) technology. The functional layers are: (1) flow layer with microchannels for transporting sample/reagent fluids, and (2) control layer with

microchannels for transmitting control signals (i.e., hydraulic or pneumatic pressure). In flow-based microfluidic biochips, microvalves on the control layer need to be connected to control pins via control channels. In application-specific and portable microfluidic devices, critical microvalves need to switch at the same time for correct functionality. Those microvalves are required to have equal or similar channel lengths to the control pin, so that the control signal can reach them simultaneously. We present a practical control-layer routing flow (PACOR) considering the critical length-matching constraint. Major features of PACOR include: (1) effective candidate Steiner tree construction and selection methods for multiple microvalves based on the deferred-merge embedding (DME) algorithm and maximum weight clique problem (MWCP) formulation, (2) minimum cost flow-based formulation for simultaneous escape routing for improved routability, and (3) minimum-length bounded routing method to detour paths for length matching. Computational simulation results show effectiveness and efficiency of PACOR with promising matching results and 100% routing completion rate.

The past decade has seen noticeable progress in computer-aided design (CAD) methods for droplet-based (digital) microfluidic biochips. However, CAD method for flow-based microfluidic biochips is still in its infancy. There are two major stages in this CAD flow: (1) control-layer design, and (2) flow-layer design. Microvalves are the critical components that closely couple these two design stages. Inferior flow-layer design solution forces valves to be placed at unfavorable positions. This makes great trouble to the following control-layer design, or even results in design failure. I.e., separate flow-layer and control-layer design lacks a global view with degraded solution quality. We have made the first attempt on flow-control co-design methodology, which integrates the two design stages for iterative adjustments with overall design improvements.

Future microfluidic biochip will be integrated with various devices, such as photodetectors and electrochemical sensors, which forms a complicated microfluidic cyber-physical system. Promising applications of such (implantable) cyber-physical microfluidic system include real-time health monitoring along with personalized preventive health care, which benefits the whole world. Microfluidic biochips are opening a door for new exciting discoveries of the unknown world. The ever-increasing integration scale of biochips drives the urgent need for CAD tools for design, modeling, and simulation.

Participants

- Mirela Alistar
Copenhagen, DK
- Krishnendu Chakrabarty
Duke University – Durham, US
- Rolf Drechsler
Universität Bremen, DE
- William H. Grover
University of California at Riverside, US
- Tsung-Yi Ho
National Tsing Hua Univ., TW
- Juinn-Dar Huang
National Chiao Tung University – Taiwan, TW
- Oliver Keszöcze
DFKI – Bremen, DE
- Bing Li
TU München, DE
- Pietro Lio'
University of Cambridge, GB
- Jan Madsen
Technical Univ. of Denmark – Lyngby, DK
- Sebastian Maerkl
EPFL – Lausanne, CH
- Paul Pop
Technical Univ. of Denmark – Lyngby, DK
- Sudip Roy
Indian Institute of Technology – Roorkee, IN
- Ulf Schlichtmann
TU München, DE
- Kwanwoo Shin
Sogang University – Seoul, KR
- Rüdiger Trojok
KIT – Karlsruher Institut für Technologie, DE
- Steve Wereley
Purdue University – West Lafayette, US
- Robert Wille
University of Bremen/DFKI, DE, and Johannes Kepler University Linz, AT
- Hailong Yao
Tsinghua University – Beijing, CN

Report from Dagstuhl Seminar 15361

Mathematical and Computational Foundations of Learning Theory

Edited by

Matthias Hein[1], Gabor Lugosi[2], and Lorenzo Rosasco[3]

1 **Universität des Saarlandes, DE,** hein@cs.uni-saarland.de
2 **UPF – Barcelona, ES,** gabor.lugosi@gmail.com
3 **MIT – Cambridge, US,** lrosasco@mit.edu

Abstract

Machine learning has become a core field in computer science. Over the last decade the statistical machine learning approach has been successfully applied in many areas such as bioinformatics, computer vision, robotics and information retrieval. The main reasons for the success of machine learning are its strong theoretical foundations and its multidisciplinary approach integrating aspects of computer science, applied mathematics, and statistics among others. The goal of the seminar was to bring together again experts from computer science, mathematics and statistics to discuss the state of the art in machine learning and identify and formulate the key challenges in learning which have to be addressed in the future. The main topics of this seminar were:

- Interplay between Optimization and Learning,
- Learning Data Representations.

Seminar August 30 to September 4, 2015 – http://www.dagstuhl.de/15361
1998 ACM Subject Classification G.1.2 Approximation, G.1.6 Optimization, G.3 Probability and Statistics, I.5 Pattern Recognition, I.2.6 Learning
Keywords and phrases learning theory, non-smooth optimization (convex and non-convex), signal processing
Digital Object Identifier 10.4230/DagRep.5.8.54

1 Executive Summary

Matthias Hein
Gabor Lugosi
Lorenzo Rosasco

Machine learning is nowadays a central field in computer science. Over the last decade the statistical learning approach has been successfully applied in many areas such as bioinformatics, computer vision, robotics and information retrieval. We believe that the main reasons for the success of machine learning are its strong theoretical foundations and its multidisciplinary approach integrating aspects of computer science, applied mathematics, and statistics among others.

Two very successful conferences titled "Mathematical Foundations of Learning Theory" in Barcelona 2004 and Paris 2006 have been inspired by this point of view on the foundations of machine learning. In 2011 the Dagstuhl seminar "Mathematical and Computational Foundations of Learning Theory" has been organized in the same spirit, bringing together

leading researchers from computer science and mathematics to discuss the state of the art and future challenges in machine learning. The 2011 Dagstuhl seminar has been the first to cover a wide range of facets of modern learning theory and has been unanimously considered a success by the participants. Since 2011 new challenges have emerged largely motivated by the availability of data-sets of unprecedented size and complexity. It is now common in many applied domains of science and technology to have datasets with thousands and even millions data-points, features and attributes/categories. For example ImageNet (http://image-net.org) is a computer vision database for object recognition including one million images of one thousands different objects, and image representations are often of the order of several tens of thousands features. Datasets of analogous complexity are customary in biology and information science (e.g. text classification). The need of analyzing and extracting information from this kind of data has posed a host of new challenges and open questions.

The second Dagstuhl seminar on "Mathematical and Computational Foundations of Learning Theory" covered broadly recent developments in the area of learning. The main focus was on two topics:

- **Interplay between Optimization and Learning**
  While statistical modeling and computational aspects have for a long time been considered separate steps in the design of learning algorithms, dealing effectively with big data requires developing new strategies where statistical and computational complexities are taken simultaneously into account. In other words, the trade-off between optimization error and generalization error has to be exploited. On the other hand it has very recently been noticed that several non-convex NP-hard learning problems (sparse recovery, compressed sensing, dictionary learning, matrix factorization etc.) can be solved efficiently and optimally (in a global sense) under conditions on the data resp. the chosen model or under the use of additional constraints.
- **Learning Data Representations**
  Data representation (e.g. the choice of kernels or features) is widely acknowledged to be the crucial step in solving learning problems. Provided with a suitable data representation, and enough labeled data, supervised algorithms, such as Support Vector Machines or Boosting, can provide good generalization performance. While data representations are often designed ad hoc for specific problems, availability of large/huge amount of unlabeled data have recently motivated the development of data driven techniques, e.g. dictionary learning, to adaptively solve the problem. Indeed, although novel tools for efficient data labeling have been developed (e.g. Amazon Mechanical Turk– http://mturk.com) most available data are unlabeled and reducing the amount of (human) supervision needed to effectively solve a task remains an important open challenge. While up-to-now the theory of supervised learning has become a mature field, an analogous theory of unsupervised and semi-supervised learning of data representation is still in its infancy and progress in the field is often assessed on a purely empirical basis.

The seminar featured a series of talks on both topics with interesting and exciting new results which lead to insights in both areas as well as a lot of discussion and interaction between the participants which for sure will manifest in several follow-up papers. Also it became obvious during the seminar that there are close connections between these two topics. Apart from these two main topics several other aspects of learning theory were discussed, leading to a quite complete picture on the current state-of-the-art in the field.

Acknowledgements. We would like to thank Dagmar Glaser and the staff at Schloss Dagstuhl for their continuous support and great hospitality which was the basis for the success of this seminar.

2 Table of Contents

3 Overview of Talks

3.1 Convex Risks, Calibrated Surrogates, Consistency, and Their Relationship with Nonparametric Estimation

Shivani Agarwal (Indian Institute of Science – Bangalore, IN)

Joint work of Agarwal, Shivani; Ramaswamy, Harish G.
Main reference H. G. Ramaswamy, S. Agarwal, "Convex calibration dimension for multiclass loss matrices," to appear in the Journal of Machine Learning Research; pre-print available as arXiv:1408.2764v2 [cs.LG], 2015.
URL http://arxiv.org/abs/1408.2764v2

In the theoretical analysis of supervised learning, the notions of PAC learning and universally Bayes consistent learning are often treated separately. We argue that classical PAC learning can essentially be viewed as a form of parametric estimation, while universally Bayes consistent learning can be viewed as a form of nonparametric estimation. A popular framework for achieving universal Bayes consistency is to minimize a (convex) calibrated surrogate risk; this is well understood for binary classification and a few selected multiclass problems, but a general understanding has remained elusive. We discuss our recent work on developing a unified framework for designing convex calibrated surrogates for general multiclass learning problems. In particular, we introduce the notion of 'convex calibration dimension' of a general multiclass loss matrix, which is the smallest number of dimensions in which one can define a convex calibrated surrogate, and give a general recipe for designing low-dimensional convex calibrated surrogates for learning problems with low-rank loss matrices. We also discuss connections between calibrated surrogates and property elicitation. In particular, we show how calibrated surrogates in supervised learning can essentially be viewed as strictly proper scoring rules for estimating certain useful properties of the conditional label distribution. These results help to shed light on how to design universally Bayes consistent algorithms for general multiclass problems, while also pointing to many open directions.

References

1 Harish G. Ramaswamy and Shivani Agarwal. *Convex calibration dimension for multiclass loss matrices*. Journal of Machine Learning Research, 2015. To appear.
2 Harish G. Ramaswamy, Shivani Agarwal and Ambuj Tewari. *Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses*. NIPS 2013.
3 Arpit Agarwal and Shivani Agarwal. *On consistent surrogate risk minimization and property elicitation*. COLT 2015.

3.2 Dictionary learning using tensor methods

Animashree Anandkumar (University of California – Irvine, US)

URL http://newport.eecs.uci.edu/anandkumar/#publications

The dictionary learning problem posits that the input data is a combination of unknown dictionary elements. Traditional methods are based on alternating minimization between the dictionary elements and coefficients. We present alternative methods based on tensor decomposition which recover the dictionary elements. These methods can consistently recover

the dictionary elements when the coefficients are independent or sufficiently uncorrelated. We also present recent extensions to the convolutional setting, where shift invariance constraints are imposed.

3.3 Optimal online prediction with quadratic loss

Peter L. Bartlett (University of California – Berkeley, US)

Joint work of Bartlett, Peter L.; Koolen, Wouter M.; Malek, Alan; Takimoto, Eiji; Warmuth, Manfred K.
Main reference P. L. Bartlett, W. M. Koolen, A. Malek, E. Takimoto, M. K. Warmuth, "Minimax Fixed-Design Linear Regression," in Proc. of the 28th Conf. on Learning Theory (COLT'15), JMLR Proceedings, Vol. 40, pp. 226–239, 2015.
URL http://jmlr.org/proceedings/papers/v40/Bartlett15.html

We consider a linear regression game in which the covariates are known in advance: at each round, the learner predicts a real value, the adversary reveals a label, and the learner incurs a squared error loss. The aim is to minimize the difference between the cumulative loss and that of the linear predictor that is best in hindsight. For a variety of constraints on the adversary's labels, we obtain an explicit expression for the minimax regret and we show that the minimax optimal strategy is linear, with a parameter choice that is reminiscent of ordinary least squares. This strategy is easy to compute and does not require knowledge of the constraint set.

We also consider the case of adversarial design, and exhibit constraint sets of covariate sequences for which the same strategy is minimax optimal.

3.4 Learning to cluster – a statistical framework for incorporating domain knowledge in clustering.

Shai Ben-David (University of Waterloo, CA)

Joint work of Ben-David, Shai; Ashtiani, Hassan
Main reference H. Ashtiani, S. Ben-David, "Representation Learning for Clustering: A Statistical Framework," in Proc. of the 31st Conf. on Uncertainty in Artificial Intelligence (UAI'15), paper ID 305, 10 pages, 2015; pre-print available as arXiv:1506.05900v1 [stat.ML], 2015.
URL http://auai.org/uai2015/proceedings/papers/305.pdf
URL http://arxiv.org/abs/1506.05900v1

Clustering is an area of huge practical relevance but rather meager theoretical foundations. The multitude of clustering algorithms (and their possible parameter settings) and the diversity of the results they may yield, call for incorporation of domain expertise in the process of selecting a clustering algorithm and setting up its parameters. I outlined recent progress made along this direction. In particular, I described a novel statistical/machine-learning approach to that challenge; a model selection algorithm that is based on interactions with the clustering user. I analyzed the statistical complexity of the proposed approach. I also mentioned some common misconceptions and potential pitfalls, aiming to stimulate discussions and highlight open questions.

References

1 Hassan Ashtiani and Shai Ben-David. *Representation Learning for Clustering: A Statistical Framework*. Proceedings of UAI 2015 and CoRR abs/1506.05900, 2015.

3.5 Is adaptive early stopping possible in statistical inverse problems?

Gilles Blanchard (Universität Potsdam, DE)

We consider a standard (mathematically idealized) setting of statistical inverse problems, taking the form of the "Gaussian sequence model" $Y_i = \lambda_i \mu_i + \varepsilon_i$, $i = 1, \ldots, D$, the random noise variables ε_i are i.i.d. Gaussian with (known) variance σ^2 , the coefficients λ_i are known, and the goal is to recover as well as possible (in the sense of squared risk) the "signal sequence" $(\mu_i)_{1 \leq i \leq D}$.

Consider the simple family of "keep or kill" estimators depending on a cutoff index k_0, that is, the corresponding estimate sequence $(\hat{\mu}_i^{(k_0)})_{1 \leq i \leq D}$ is simply equal to $\lambda_i^{-1} Y_i$ for $i < k_0$ and 0 for $k_0 \leq i \leq D$. The question of *adaptivity* is the following: is it possible to choose $\hat{k}_0$ from the data only, in such a way that the performance obtained is comparable (whithin a multiplicative constant) to the best possible deterministic, a priori choice of k_0 minimising the average squared risk (usually called "oracle", since it depends on the unknown signal)?

There exist a number of well-known methods achieving oracle adaptivity, such as penalization or Lepski's method. However, they have in common that the estimators for *all* values of k_0 have to be computed first and compared to each other in some way. Contrast this to an "early stopping" approach where we would like to compute iteratively the estimators for $k_0 = 1, 2, \ldots$ and have to decide to stop at some point $\hat{k}_0$ without being allowed to compute the other estimators. Is oracle adaptivity possible then? This question is motivated by settings where computing estimators for larger k_0 requires more computational cost; furthermore some form of early stopping is most often used in practice.

After careful mathematical formalization of the problem, our first result is that, if one must base the early stopping decision at index k_0 on the sole information of $Y_i, i \leq k_0$, then adaptive early stopping is not possible in general. A more realistic scenario is when we are additionally allowed to use the information of the *residual* $\sum_{i=k_0+1}^{D} Y_i^2$ to decide to stop at k_0 (or not). In that case, partial oracle adaptation is possible, essentially when the oracle stopping time k_0^* is larger in order than $\sqrt{D}$ (remember D is the maximum considered dimension). This adaptive stopping can be achieved by a simple "discrepancy principle" commanding to stop when the residual becomes smaller than $D\sigma^2$, a type of rule which is often used in practice. We establish lower and upper bounds, in particular showing that if the oracle k_0^* is of order strictly smaller than $\sqrt{D}$, oracle adaptation is *not* possible in general.

3.6 Adaptive tail index estimation

Stéphane Boucheron (Paris Diderot University, FR)

Joint work of Stéphane Boucheron, Maud Thomas
Main reference S. Boucheron, M. Thomas, "Tail index estimation, concentration and adaptivity," arXiv:1503.05077v3 [math.ST], 2015.
URL http://arxiv.org/abs/1503.05077v3

Assume data $X_1, \ldots, X_n$ are collected from a univariate distribution F and we want to estimate $\overline{F}(x) = 1 - F(x)$ where $x > \overline{F}(\max(X_1, \ldots, X_n)$ or estimate a quantile of order $1 - 1/t$ for $t > n$. In order the face this challenge with a reasonable of possibility of success, a

tail regularity assumption is necessary. In the so-called heavy tail domains, this assumption reads as: for all $x > 0$, $\lim_{t\to\infty} \overline{F}(tx)/\overline{F}(t) = x^{-1/\gamma}$ for some $\gamma > 0$ which is called the tail (or extreme value) index. In words, $\overline{F}$ is assumed to be regularly varying with index $-1/\gamma$. Estimating γ from a sample is called the tail index estimation problem (see [4] for a presentation of Extreme Value Theory). Many tail index estimators (Hill, Pickands, Moments, ...) consist of computing statistics from the k largest order statistics. Practitioners face an estimator selection problem: picking k so as to achieve a good trade-off between variance (large values of k) and bias (small values of k). We present an adaptive version of the Hill estimator based on Lespki's model selection method (which has been used in learning theory in order to achieve adaptivity in classification, see [2, 3]). This simple data-driven index selection method is shown to satisfy an kind of oracle inequality and is checked to achieve the lower risk bound recently derived by [1]. In order to establish the (pseudo)-oracle inequality, we derive non-asymptotic variance bounds and concentration inequalities for Hill estimators. These concentration inequalities are derived from Talagrand's concentration inequality for smooth functions of independent exponentially distributed random variables combined with three tools of Extreme Value Theory: the quantile transform, Karamata's representation of slowly varying functions, and Rényi's characterisation for the order statistics of exponential samples.

References

1 Carpentier, Alexandra and Kim, Arlene K.H. *Adaptive and minimax optimal estimation of the tail coefficient*. arXiv:1309.2585v1, 2013.
2 Tsybakov, Alexandre B. *Optimal aggregation of classifiers in statistical learning*. Annals of Statistics 32, 135–166, 2004.
3 Boucheron, Stéphane and Bousquet, Olivier and Lugosi, Gábor. *Theory of Classification: a Survey of Some Recent Advances*. ESAIM: Probability and Statistics 9, 329–375, 2005.
4 Beirlant, Jan and Goegebeur, Yuri and Segers, Johan and Teugels, Jozef. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

3.7 Multi-scale exploration of convex functions and bandit convex optimization

Sébastien Bubeck (Microsoft Research – Redmond, US)

Joint work of Bubeck, Sébastien; Eldan, Ronen
Main reference S. Bubeck, R. Eldan, "Multi-scale exploration of convex functions and bandit convex optimization," arXiv:1507.06580v1 [math.MG], 2015.
URL http://arxiv.org/abs/1507.06580v1

We construct a new map from a convex function to a distribution on its domain, with the property that this distribution is a multi-scale exploration of the function. We use this map to solve a decade-old open problem in adversarial bandit convex optimization by showing that the minimax regret for this problem is $\tilde{O}(poly(n)\sqrt{T})$, where n is the dimension and T the number of rounds. This bound is obtained by studying the dual Bayesian maximin regret via the information ratio analysis of Russo and Van Roy, and then using the multi-scale exploration to solve the Bayesian problem.

3.8 Information theory of algorithms

Joachim M. Buhmann (ETH Zürich, CH)

Main reference J. M. Buhmann, "SIMBAD: Emergence of Pattern Similarity, Similarity-Based Pattern Analysis and Recognition," in M. Pelillo (ed.), "Similarity-Based Pattern Analysis and Recognition – Part I", pp. 45–64, Springer, 2013.
URL http://dx.doi.org/10.1007/978-1-4471-5628-4_3

Algorithms map input spaces to output spaces where inputs are possibly affected by fluctuations. Beside run time and memory consumption, an algorithm might be characterized by its sensitivity to the signal in the input and its robustness to input fluctuations. The achievable precision of an algorithm, i.e., the attainable resolution in output space, is determined by its capability to extract predictive information in the input relative to its output. I will present an information theoretic framework for algorithm analysis where an algorithm is characterized as computational evolution of a (possibly contracting) posterior distribution over the output space. The tradeoff between precision and stability is controlled by an input sensitive generalization capacity (GC). GC measures how much the posteriors on two different problem instances agree despite the noise in the input. Thereby, GC objectively ranks different algorithms for the same data processing task based on the bit rate of their respective capacities. Information theoretic algorithm selection is demonstrated for minimum spanning tree algorithms and for greedy MaxCut algorithms. The method can rank centroid based and spectral clustering methods, e.g. k-means, pairwise clustering, normalized cut, adaptive ratio cut and dominant set clustering.

3.9 Fast algorithms and (other) minimax optimal algorithms for mixed regression

Constantine Caramanis (Univ. of Texas at Austin, US)

Joint work of Caramanis, Constantine; Chen, Yudong; Yi, Xinyang
Main reference Y. Chen, X. Yi, C. Caramanis, "A Convex Formulation for Mixed Regression with Two Components: Minimax Optimal Rates," in Proc. of the 27th Conf. on Learning Theory (COLT'14), JMLR Proceedings, Vol. 35, pp. 560–604, 2014; pre-print available as arXiv:1312.7006v2 [stat.ML], 2015.
URL http://jmlr.org/proceedings/papers/v35/chen14.html
URL http://arxiv.org/abs/1312.7006v2

Mixture models represent the superposition of statistical processes, and are natural in machine learning and statistics. In mixed regression, the relationship between input and output is given by one of possibly several different (noisy) linear functions. Thus the solution encodes a combinatorial selection problem, and hence computing it is difficult in the worst case. Even in the average case, little is known in the realm of efficient algorithms with strong statistical guarantees.

We give general conditions for linear convergence of an EM-like (and hence fast) algorithm for latent-variable problems in high dimensions, and show this implies that for sparse (or low-rank) mixed regression, EM converges linearly, in a neighborhood of the optimal solution, in the high-SNR regime. For the low-SNR regime, we show that a new behavior emerges. Here, we give a convex optimization formulation that provably recovers the true solution, and we provide upper bounds on the recovery errors for both arbitrary noise and stochastic

noise settings. We also give matching minimax lower bounds, showing that our algorithm is information-theoretically optimal.

Our results represent what is, as far as we know, the only tractable algorithm guaranteeing successful recovery with tight bounds on recovery errors and sample complexity.

References

1 Yudong Chen, Xinyang Yi, Constantine Caramanis. *A Convex Formulation for Mixed Regression with Two Components: Minimax Optimal Rates*. JMLR W&CP, 35:560–604, 2014
2 Xinyang Yi, Constantine Caramanis. *Regularized EM Algorithms: A Unified Framework and Statistical Guarantees*. To appear at NIPS 2015

3.10 Sparse and spurious: dictionary learning with noise and outliers

Rémi Gribonval (INRIA Rennes – Bretagne Atlantique, FR)

Joint work of Gribonval, Rémi; Jenatton, Rodolphe; Bach, Francis; Kleinsteuber, Martin; Seibert, Matthias;
Main reference R. Gribonval, R. Jenatton, F. R. Bach, "Sparse and Spurious: Dictionary Learning With Noise and Outliers," IEEE Transactions on Information Theory, 61(11):6298–6319, 2015: pre-print available as arXiv:1407.5155v4 [cs.LG], 2015.
URL http://dx.doi.org/10.1109/TIT.2015.2472522
URL http://arxiv.org/abs/1407.5155v4

In this talk I draw a panorama of dictionary learning for low-dimensional modeling. After reviewing the basic empirical principles of dictionary learning and related matrix factorizations such as PCA, K-means and NMF, I discuss techniques to learn dictionaries with controlled computational efficiency, as well as a series of recent theoretical results establishing the statistical significance of learned dictionaries even in the presence of noise and outliers.

References

1 Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, Matthias Seibert. *Sample Complexity of Dictionary Learning and other Matrix Factorizations*. IEEE Transactions on Information Theory, 2015
2 Rémi Gribonval, Rodolphe Jenatton, Francis Bach. *Sparse and spurious: dictionary learning with noise and outliers*. IEEE Transactions on Information Theory, 2015

3.11 Empirical portfolio selections and a problem on aggregation

László Györfi (Budapest University of Technology & Economics, HU)

Main reference L. Györfi, Gy. Ottucsák, A. Urbán, "Empirical log-optimal portfolio selections: a survey," in L. Györfi, G. Ottucsák, H. Walk (eds.), "Machine Learning for Financial Engineering," pp. 79–116, Imperial College Press, 2012.
URL http://www.worldscientific.com/worldscibooks/10.1142/p818

This talk provides a survey of discrete time, multi period, equential investment strategies for financial markets. Under memoryless assumption on the underlying process generating the asset prices the Best Constantly Rebalanced Portfolio is studied, called log-optimal portfolio, which achieves the maximal asymptotic average growth rate. For generalized dynamic portfolio selection, when asset prices are generated by a stationary and ergodic process,

growth optimal empirical strategies are shown, where some principles of nonparametric regression estimation and of machine learning aggregation are applied. The empirical performance of the methods is illustrated for NYSE data. An open problem is presented, too, which means that the consistency has been proved if the learning parameter for the aggregation is between 0 and 1, while the empirical results are better if the learning parameter is larger than 1. The problem is to extend the consistency to this case.

3.12 Train faster, generalize better: Stability of stochastic gradient descent

Moritz Hardt (Google Research – Mountain View, US)

We show that any model trained by a stochastic gradient method with few iterations has vanishing generalization error. We prove this by showing the method is algorithmically stable in the sense of Bousquet and Elisseeff. Our analysis only employs elementary tools from convex and continuous optimization. Our results apply to both convex and non-convex optimization under standard Lipschitz and smoothness assumptions.

Applying our results to the convex case, we provide new explanations for why multiple epochs of stochastic gradient descent generalize well in practice. In the nonconvex case, we provide a new interpretation of common practices in neural networks, and provide a formal rationale for stability-promoting mechanisms in training large, deep models. Conceptually, our findings underscore the importance of reducing training time beyond its obvious benefit.

3.13 Robust Regression via Hard Thresholding

Prateek Jain (Microsoft Research India – Bangalore, IN)

Joint work of Jain, Prateek; Bhatia Kush; Kar Purushottam
Main reference K. Bhatia, P. Jain, P. Kar, "Robust Regression via Hard Thresholding," to appear in Proc. of the 29th Annual Conf. on Neural Information Processing Systems (NIPS'15): pre-print available as arXiv:1506.02428v1 [cs.LG], 2105.
URL http://arxiv.org/abs/1506.02428v1

In this talk, we will discuss the problem of Robust Least Squares Regression (RLSR) where several response variables can be adversarially corrupted. More specifically, for a data matrix $X \in R^{p \times n}$ and an underlying model w^*, the response vector is generated as $y = X'w^* + b$ where $b \in R^n$ is the corruption vector supported over at most $C\,n$ coordinates. Existing exact recovery results for RLSR focus solely on L1-penalty based convex formulations and impose relatively strict model assumptions such as requiring the corruptions b to be selected independently of X. In this talk, we will focus on a simple hard-thresholding algorithm that we call TORRENT which, under mild conditions on X, can recover w^* exactly even if b corrupts the response variables in an adversarial manner, i.e. both the support and entries of b are selected adversarially after observing X and w^*. We will also discuss certain extensions of TORRENT that can scale efficiently to large scale problems, such as high dimensional sparse recovery. We will present empirical results that show that TORRENT,

and more so its extensions, offer significantly faster recovery than the state-of-the-art L1 solvers. For instance, even on moderate-sized datasets (with $p = 50K$) with around 40% corrupted responses, a variant of our proposed method called TORRENT-HYB is more than 20x faster than the best L1 solver.

See http://arxiv.org/abs/1506.02428 for more details.

3.14 Optimizing decomposable submodular functions

Stefanie Jegelka (MIT – Cambridge, US)

Submodular functions capture a spectrum of discrete problems in machine learning, signal processing and computer vision. In these areas, practical algorithms are a major concern that motivates to exploit structure in addition to submodularity. A simple example of such a structure are functions that decompose as a sum of "simple" submodular functions. For this setting, several algorithms arise from relations between submodularity and convexity. In particular, this talk will focus on a class of algorithms that solve submodular minimization as a best approximation problem. These algorithms are easy to use and to parallelize, and solve both a convex relaxation and the original discrete problem. We observe that the algorithms work well in practice, and analyze their convergence properties.

References

1 S. Jegelka, F. Bach, S. Sra. *Reflection methods for user-friendly submodular optimization*. NIPS 2013
2 R. Nishihara, S. Jegelka, M.I. Jordan. *On the linear convergence rate of decomposable submodular function minimization*. NIPS 2014

3.15 Matrix factorization with binary components – uniqueness in a randomized model

Felix Krahmer (TU München, DE)

Joint work of Hein, Matthias; James, David; Krahmer, Felix

Motivated by an application in computational biology, we consider low-rank matrix factorization with $\{0, 1\}$-constraints on the first of the factors and optionally convex constraints on the second one. Despite apparent intractability, it has been shown by Hein et al. [1] that one can provably recover the underlying factorization, provided there exists a unique solution. We conjecture that by choosing a sparse Bernoulli random model for the binary factor, there will be a unique solution with high probability. Due to limited applicability of Littlewood-Offord inequalities, previous results do not generalize. We present partial progress for limited rank.

References

1 M. Slawski, M. Hein, and P. Lutsik, *Matrix Factorization with Binary Components*. NIPS 2013

3.16 Variational Inference in Probabilistic Submodular Models

Andreas Krause (ETH Zürich, CH)

Joint work of Djolonga, Josip; Krause, Andreas
Main reference J. Djolonga, A. Krause, "From MAP to Marginals: Variational Inference in Bayesian Submodular Models," in Proc. of the 28th Annual Conf. on Advances in Neural Information Processing Systems (NIPS'14), pp. 244–252, 2014.
URL http://papers.nips.cc/paper/5492-from-map-to-marginals-variational-inference-in-bayesian-submodular-models

As a discrete analogue of convexity, submodularity has profound implications for optimization. In recent years, submodular optimization has found many new applications, such as in machine learning and network analysis. These include active learning, dictionary learning, data summarization, influence maximization and network structure inference. In this talk, I will present our recent work on quantifying uncertainty in submodular optimization. In particular, we carry out the first systematic investigation of inference and learning in probabilistic submodular models (PSMs). These are probabilistic models defined through submodular functions – log-sub/supermodular distributions – generalizing regular binary Markov Random Fields and Determinantal Point Processes. They express natural notions such as attractiveness and repulsion and allow to capture long-range, high-order dependencies among the variables. I will present our recently discovered variational approach towards inference in general PSMs based on sub- and supergradients. We obtain both lower and upper bounds on the log- partition function, which enables computing probability intervals for marginals, conditionals and marginal likelihoods. We also obtain fully factorized approximate posteriors, at essentially the same computational cost as ordinary submodular optimization. Our framework results in convex problems for optimizing over differentials of submodular functions, which we show how to optimally solve. Our approximation is exact at the mode (for log-supermodular distributions), and we provide bounds on the approximation quality of the log-partition function with respect to the curvature of the function. We further establish natural relations between our variational approach and the classical mean-field method from statistical physics. Exploiting additive structure in the objective leads to highly scalable, parallelizable message passing algorithms. We empirically demonstrate the accuracy of our inference scheme on several PSMs arising in computer vision and network analysis.

3.17 Learning Representations from Incomplete Data

Robert D. Nowak (University of Wisconsin – Madison, US)

Joint work of Nowak, Robert D; Pimentel, Daniel; Boston, Nigel
Main reference D. L. Pimentel-Alarcón, N. Boston, R. D. Nowak, "A Characterization of Deterministic Sampling Patterns for Low-Rank Matrix Completion," arXiv:1503.02596v2 [stat.ML] , 2015.
URL http://arxiv.org/abs/1503.02596v2

Low-rank matrix completion (LRMC) problems arise in a wide variety of applications. Previous theory mainly provides conditions for completion under missing-at-random samplings. This talk presents deterministic conditions for completion. An incomplete $d \times N$ matrix is finitely rank-r completable if there are at most finitely many rank-r matrices that agree with all its observed entries. Finite completability is the tipping point in LRMC, as a few additional samples of a finitely completable matrix guarantee its unique completability.

The main contribution the talk is a characterization of finitely completable observation sets. We use this characterization to derive sufficient deterministic sampling conditions for unique completability. We also show that under uniform random sampling schemes, these conditions are satisfied with high probability if $O(\max\{r, \log d\})$ entries per column are observed. Extensions of these results to subspace clustering with missing data are also given.

Further details can be found in the following papers: arXiv:1503.02596, arXiv:1410.0633

3.18 Tight convex relaxations for sparse matrix factorization

Guillaume Obozinski (ENPC – Marne-la-Vallée, FR)

Joint work of Richard, Emile; Vert, Jean-Philippe

In this talk, I will consider statistical learning problems in which the parameter is a matrix which is the sum of a small number of sparse rank one (non-orthogonal) factors, and which can be viewed as generalizations of the sparse PCA problem with multiple factors. Based on an assumption that the sparsity of the factors is fixed and known, I will present a matrix norm which provides an tight although NP-hard convex relaxation of the learning problem. I will discuss the sample complexity of learning the matrix in the rank one case and show that considering a computationally more expensive convex relaxation leads to an improvement of the sample complexity by an order of magnitude as compared with the usual convex regularization considered, like combinations of the L1-norm and the trace norm. I will then describe an algorithm, relying on a rank-one sparse PCA oracle to solve the convex problems considered and illustrate that, in practice, when state-of-the-art heuristic algorithms for rank one sparse PCA are used as surrogates for the oracle, our algorithm outperforms other existing methods.

3.19 Active Regression

Sivan Sabato (Ben Gurion University – Beer Sheva, IL)

Joint work of Rémi Munos
Main reference S. Sabato, R. Munos, "Active Regression by Stratification," in Proc. of the 28th Annual Conf. on Advances in Neural Information Processing Systems (NIPS'14), pp. 269–477, 2014.
URL http://papers.nips.cc/paper/5468-active-regression-by-stratification

We propose a new active learning algorithm for parametric linear regression with random design. We provide finite sample convergence guarantees for general distributions in the misspecified model. This is the first active learner for this setting that provably can improve over passive learning. Unlike other learning settings (such as classification), in regression the passive learning rate of $O(1/\epsilon)$ cannot in general be improved upon. Nonetheless, the so-called 'constant' in the rate of convergence, which is characterized by a distribution- dependent *risk*, can be improved in many cases. For a given distribution, achieving the optimal risk requires prior knowledge of the distribution. Following the stratification technique advocated in Monte-Carlo function integration, our active learner approaches the optimal risk using piecewise constant approximations.

Sivan Sabato is supported by the Lynne and William Frankel Center for Computer Science.

3.20 Dictionary learning – fast and dirty

Karin Schnass (Universität Innsbruck, AT)

Main reference K. Schnass, "Convergence radius and sample complexity of ITKM algorithms for dictionary learning," arXiv:1503.07027v2 [cs.LG], 2015.
URL http://arxiv.org/abs/1503.07027v2

In this talk we give a short introduction to fast dictionary learning algorithms with local convergence guarantees. Using the classic optimization principle underlying K-SVD as starting point we motivate a response maximization principle and and the associated algorithm ITKM (Iterative Thresholding and K Means). We then progress to a variant using residual means ITKrM (Iterative Thresholding and K residual Means), which can be seen as hybrid between K-SVD and ITKrM and as such inherits the best of both worlds. Experimental global convergence from K-SVD, and computational efficiency, sequentiality/parallelizability and local convergence guarantees under low sample complexity from ITKM.

3.21 Variational approach to consistency of clustering of point clouds

Dejan Slepcev (Carnegie Mellon University, US)

Joint work of Garcia Trillos, Nicolas; Laurent, Thomas; von Brecht, James; Bresson, Xavier; Slepcev, Dejan
Main reference N. Garcia Trillos, D. Slepcev, J. van Brecht, T. Laurent, X. Bresson, "Consistency of Cheeger and Ratio Graph Cuts," arXiv:1411.6590v1 [stat.ML], 2014.
URL http://arxiv.org/abs/1411.6590v1

The talk discussed variational problems arising in machine learning and their consistency as the number of data points goes to infinity. Consider point clouds obtained as random samples of an underlying "ground-truth" measure on a Euclidean domain. Graph representing the point cloud is obtained by assigning weights to edges based on the distance between the points. We discussed approaches to clustering based on minimizing objective functionals defined on these graphs. We focused is on functionals based on graph cuts like the Cheeger and ratio cuts. We showed that minimizers of the these cuts converge as the sample size increases to a minimizer of a corresponding continuum cut (which partitions the ground truth measure). A setup based on Gamma-convergence and optimal transportation to study such questions was introduced. Sharp conditions on how the connectivity radius can be scaled with respect to the number of sample points for the consistency to hold were obtained.

3.22 Optimization, Regularization and Generalization in Multilayer Networks

Nathan Srebro (TTIC – Chicago, US)

Joint work of Srebro, Nathan; Neyshabur, Behnam; Tomioka, Ryota; Salakhutdinov, Russ

What is it that enables learning with multi-layer networks? What causes the network to generalize well? What makes it possible to optimize the error, despite the problem being

hard in the worst case? In this talk I will attempt to address these questions and relate between them, highlighting the important role of optimization in deep learning. I will then use the insight to suggest studying novel optimization methods, and will present Path-SGD, a novel optimization approach for multi-layer RELU networks that yields better optimization and better generalization.

3.23 Oracle inequalities for network models and sparse graphon estimation

Alexandre Tsybakov (UPMC – Paris, FR)

Joint work of Klopp, Olga; Tsybakov, Alexandre B.; Verzelen, Nicolas
Main reference O. Klopp, A. B. Tsybakov, N. Verzelen, "Oracle inequalities for network models and sparse graphon estimation," arXiv:1507.04118v1 [math.ST], 2015.
URL http://arxiv.org/abs/1507.04118v1

Inhomogeneous random graph models encompass many network models such as stochastic block models and latent position models. In this paper, we study two estimators: the ordinary block constant least squares estimator, and its restricted version. We show that they satisfy oracle inequalities with respect to the block constant oracle. As a consequence, we derive optimal rates of estimation of the probability matrix. Our results cover the important setting of sparse networks. Nonparametric rates for graphon estimation in the L_2 norm are also derived when the probability matrix is sampled according to a graphon model. The results shed light on the differences between estimation under the empirical loss (the probability matrix estimation) and under the integrated loss (the graphon estimation).

3.24 Learning Economic Parameters from Revealed Preferences

Ruth Urner (MPI für Intelligente Systeme – Tübingen, DE)

Joint work of Balcan, Maria-Florina; Daniely, Amit; Mehta, Ruta; Urner, Ruth; Vazirani, Vijay V.
Main reference M.-F. Balcan, A. Daniely, R. Mehta, R. Urner, V. V. Vazirani, "Learning Economic Parameters from Revealed Preferences," in Proc. of the 10th Int'l Conf. on Web and Internet Economics (WINE'14), LNCS, Vol. 8877, pp. 338–353, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-13129-0_28

A recent line of work, starting with Beigman and Vohra and Zadimoghaddam and Roth, has addressed the problem of learning a utility function from revealed preference data. The goal here is to make use of past data describing the purchases of a utility maximizing agent when faced with certain prices and budget constraints in order to produce a hypothesis function that can accurately forecast the future behavior of the agent.

In this work we advance this line of work by providing sample complexity guarantees and efficient algorithms for a number of important classes. By drawing a connection to recent advances in multi-class learning, we provide a computationally efficient algorithm with tight sample complexity guarantees ($\Theta(d/\epsilon)$ for the case of d goods) for learning linear utility functions under a linear price model. This solves an open question in Zadimoghaddam and Roth. Our technique yields numerous generalizations including the ability to learn other

well-studied classes of utility functions, to deal with a misspecified model, and with non-linear prices.

References

1 Maria-Florina Balcan, Amit Daniely, Ruta Mehta, Ruth Urner and Vijay V. Vazirani. *Learning Economic Parameters from Revealed Preferences*. Web and Internet Economics – 10th International Conference (WINE) 2014, Beijing, China, December 14–17, 2014. Proceedings.

3.25 Stochastic Forward-Backward Splitting

Silvia Villa (Italian Institute of Technology – Genova, IT)

Joint work of Rosasco, Lorenzo; Vu, Cong Bang; Villa, Silvia
Main reference L. Rosasco, S. Villa, B. C. Vu, "Convergence of stochastic proximal gradient algorithm," arXiv:1403.5074v3 [math.OC], 2014.
URL http://arxiv.org/abs/1403.5074v3

I analyzed the convergence of a novel stochastic forward-backward splitting algorithm for solving monotone inclusions given by the sum of a maximal monotone operator and a single-valued maximal monotone cocoercive operator. This latter framework has a number of interesting special cases, including variational inequalities and convex minimization problems, while stochastic approaches are practically relevant to account for perturbations in the data. The algorithm I discussed is a stochastic extension of the classical deterministic forward-backward method, and is obtained considering the composition of the resolvent of the maximal monotone operator with a forward step based on a stochastic estimate of the single-valued operator.

The talk was based on the following papers:

References

1 L. Rosasco, S. Villa, and B. C. Vu. *Convergence of stochastic proximal gradient algorithm*. arxiv 1403.5074
2 L. Rosasco, S. Villa, and B. C. Vu. *Stochastic forward-backward splitting for monotone inclusions*. arxiv 1403.7999
3 L. Rosasco, S. Villa, and B. C. Vu. *A stochastic inertial forward-backward splitting algorithm for multivariate monotone inclusions*. arXiv:1507.00848

3.26 Finding global k-means clustering solutions

Rachel Ward (University of Texas – Austin, US)

K-means clustering aims to partition a set of n points into k clusters in such a way that each observation belongs to the cluster with the nearest mean, and such that the sum of squared distances from each point to its nearest mean is minimal. In general, this is a hard optimization problem, requiring an exhaustive search over all possible partitions of the data into k clusters in order to find the optimal clustering. At the same time, fast heuristic

algorithms for the k-means optimization problem are often applied in many data processing applications, despite having few guarantees on the clusters they produce. In this talk, we will introduce a semidefinite programming relaxation of the k-means optimization problem, along with geometric conditions on a set of data such that the algorithm is guaranteed to find the optimal k-means clustering for the data. For points drawn randomly within separated balls, the important quantities are the distances between the centers of the balls compared to the relative densities of points within them, and at sufficient density, the SDP relaxation is guaranteed to resolve such clusters at arbitrarily small separation distance. We will also discuss certain convex relaxations and recovery guarantees for another geometric clustering objective, k-median clustering. We will conclude by discussing several open questions related to this work.

3.27 Symmetric and Asymmetric k-Center Clustering under Stability

Colin White (Carnegie Mellon University, US)

Joint work of Balcan, Maria-Florina; Haghtalab, Nika; White, Colin
Main reference M.-F. Balcan, N. Haghtalab, C. White, "Symmetric and Asymmetric k-center Clustering under Stability," arXiv:1505.03924v2 [cs.DS], 2015.
URL http://arxiv.org/abs/1505.03924v2

In this work, we take a beyond the worst case approach to asymmetric and symmetric k-center problems under two very natural input stability (promise) conditions. We consider both the α-perturbation resilience notion of Bilu and Linial [BL12], which states that the optimal solution does not change under any α-factor perturbation to the input distances, and the (α,ϵ)-approximation stability notion of Balcan et al. [BBG09], which states that any α-approximation to the cost of the optimal solution should be ϵ-close in the solution space (i.e., the partitioning) to the optimal solution. We show that by merely assuming 3-perturbation resilience or $(2, 0)$-approximation stability, the exact solution for the asymmetric k-center problem can be found in polynomial time. To our knowledge, this is the first problem that is hard to approximate to any constant factor in the worst case, yet can be optimally solved in polynomial time under perturbation resilience for a constant value of α. In the case of 2-approximation stability, we prove our result is tight by showing k-center under (2-ϵ)-approximation stability is hard unless $NP = RP$. For the case of symmetric k-center, we give an efficient algorithm to cluster 2-perturbation resilient instances. Our results illustrate a surprising relation between symmetric and asymmetric k-center instances under these stability conditions. Unlike approximation ratio, for which symmetric k-center is easily solved to a factor of 2 but asymmetric k-center cannot be approximated to any constant factor, both symmetric and asymmetric k-center can be solved optimally under resilience to small constant-factor perturbations.

3.28 A Dynamic Approach to Variable Selection and Sparse Recovery: Differential Inclusions with Early Stopping

Yuan Yao (Peking University, CN)

Sparse signal recovery from linear noisy measurements has been a classical topic in compressed sensing and high dimensional statistics. There has been a large volume of literature around l1-regularization or LASSO approach and it is well-known that the convex relaxation in LASSO leads to biased solutions. So in practice, people compute LASSO regularization paths for model selection, followed by a subset least square to remove the bias. Here we discuss an alternative approach to sparse recovery via differential equations with inclusion constraints, which we call Bregman ISS (Inverse Scale Space) or Linearized Bregman ISS. We shall see that the new approach has great advantages over LASSO in its algorithmic simplicity and estimate quality. Its dynamics naturally induces a solution path for regularization and the points on the paths can be unbiased or less biased than LASSO. We show that under nearly the same conditions for LASSO's sign consistency, there exists a bias-free and sign-consistent point on the solution paths, where early stopping is crucial for regularization.

3.29 Minimum Error Entropy and Related Problems

Ding-Xuan Zhou (City University – Hong Kong, HK)

Minimum error entropy principle has been widely used in the community of signal processing and is closely related to kernel methods in learning theory. Its idea is to seek as much information as possible from data by minimizing various entropies of the error random variable. A minimum error entropy method takes moments of all orders into consideration and may perform well in dealing with heavy-tailed noise. Compared with its practical developments within the last decade, its rigorous theoretical consistency analysis is unknown. This talk demonstrates some rigorous consistency analysis of the minimum error entropy principle in the framework of regression. Some new methods arise from the study and might be used for investigating other related problems: Fourier analysis of the generalization error associated with pairwise loss functions, minimax rates of convergence achieved by the least squares regularization scheme, and the choice of step sizes for online or gradient descent algorithms.

Participants

- Shivani Agarwal
Indian Institute of Science – Bangalore, IN
- Animashree Anandkumar
Univ. of California – Irvine, US
- Peter L. Bartlett
University of California – Berkeley, US
- Shai Ben-David
University of Waterloo, CA
- Gilles Blanchard
Universität Potsdam, DE
- Stephane Boucheron
Paris Diderot University, FR
- Sebastien Bubeck
Microsoft Res. – Redmond, US
- Joachim M. Buhmann
ETH Zürich, CH
- Constantine Caramanis
University of Texas at Austin, US
- Sou-Cheng Choi
NORC – Chicago, US
- Luc Devroye
McGill Univ. – Montreal, CA
- Jack Fitzsimons
University of Oxford, GB
- Antoine Gautier
Universität des Saarlandes, DE
- Remi Gribonval
INRIA Rennes – Bretagne Atlantique, FR
- László Györfi
Budapest University of Technology & Economics, HU
- Moritz Hardt
Google Research – Mountain View, US
- Matthias Hein
Universität des Saarlandes, DE
- Prateek Jain
Microsoft Research India – Bangalore, IN
- Stefanie Jegelka
MIT – Cambridge, US
- Felix Krahmer
TU München, DE
- Andreas Krause
ETH Zürich, CH
- Lek-Heng Lim
University of Chicago, US
- Gabor Lugosi
UPF – Barcelona, ES
- Robert D. Nowak
University of Wisconsin – Madison, US
- Guillaume Obozinski
ENPC – Marne-la-Vallée, FR
- Duy Khanh Pham
Ho Chi Minh City University of Pedagogy, VN
- Lorenzo Rosasco
MIT – Cambridge, US
- Alessandro Rudi
MIT – Cambridge, US
- Sivan Sabato
Ben Gurion University – Beer Sheva, IL
- Karin Schnass
Universität Innsbruck, AT
- Dejan Slepcev
Carnegie Mellon University, US
- Nathan Srebro
TTIC – Chicago, US
- Yannik Stein
FU Berlin, DE
- Alexandre Tsybakov
UPMC – Paris, FR
- Ruth Urner
MPI für Intelligente Systeme – Tübingen, DE
- Silvia Villa
Italian Institute of Technology – Genova, IT
- Rachel Ward
University of Texas – Austin, US
- Colin White
Carnegie Mellon University, US
- Robert C. Williamson
Australian National Univ., AU
- Yuan Yao
Peking University, CN
- Ding-Xuan Zhou
City University – Hong Kong, HK

Report from Dagstuhl Perspectives Workshop 15362

Present and Future of Formal Argumentation

Edited by

Dov M. Gabbay[1], Massimiliano Giacomin[2], Beishui Liao[3], and Leendert van der Torre[4]

1 **King's College London, GB,** dov.gabbay@kcl.ac.uk
2 **University of Brescia, IT,** massimiliano.giacomin@unibs.it
3 **Zhejiang University, CN,** baiseliao@zju.edu.cn
4 **University of Luxembourg, LU,** leon.vandertorre@uni.lu

Abstract

This report documents the program and the outcomes of Dagstuhl Perspectives Workshop 15362 "Present and Future of Formal Argumentation". The goal of this Dagstuhl Perspectives Workshop was to gather the world leading experts in formal argumentation in order to develop a SWOT (Strength, Weaknesses, Opportunities, Threats) analysis of the current state of the research in this field and to draw accordingly some strategic lines to ensure its successful development in the future. A critical survey of the field has been carried out through individual presentations and collective discussions. Moreover, working group activity lead to identify several open problems in argumentation.

Perspectives Workshop August 30 to September 4, 2015 – http://www.dagstuhl.de/15362
1998 ACM Subject Classification I.2.4 Knowledge Representation Formalisms and Methods
Keywords and phrases Argumentation, Non-monotonic Logic, Multi-Agent Systems
Digital Object Identifier 10.4230/DagRep.5.8.74

1 Executive Summary

Dov M. Gabbay
Massimiliano Giacomin
Beishui Liao
Leendert van der Torre

Diverse kinds of reasoning and dialogue activities can be captured by argumentation models in a formal and still quite intuitive way, thus enabling the integration of different specific techniques and the development of applications humans can trust. Formal argumentation lays on solid bases, such as extensively studied theoretical models at different levels of abstraction, efficient implementations of these models, as well as a variety of experimental studies in several application fields. In order to be able to convert the opportunities of the present into actual results in the future, the formal argumentation research community needs however to reflect about the current assets and weaknesses of the field and to identify suitable strategies to leverage the former and to tackle the latter. As an example, the definition of standard modeling languages and of reference sets of benchmark problems are still in their infancy, reference texts for newcomers are missing, the study of methodological guidelines for the use of theoretical models in actual applications is a largely open research issue.

Present and Future of Formal Argumentation, *Dagstuhl Reports*, Vol. 5, Issue 8, pp. 74–89
Editors: Dov M. Gabbay, Massimiliano Giacomin, Beishui Liao, and Leendert van der Torre
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The goal of this Dagstuhl Perspectives Workshop was to gather the world leading experts in formal argumentation in order to develop a SWOT (Strength, Weaknesses, Opportunities, Threats) analysis of the current state of the research in this field and to draw accordingly some strategic lines to ensure its successful development in the future.

The Perspectives Workshop was held between August 30 to September 4, 2015, with 22 participants from 10 countries. With the aim of developing a critical survey of the field for the argumentation community and for potential newcomers, the organizers agreed to assemble a handbook of formal argumentation, and encouraged participants to present their view on different topics in the area. Besides individual presentations, the program included collective discussions on general issues arising from individual presentations, as well as working groups.

Individual presentations concerned introductory overviews, logical problems and requirements for formal argumentation, specific formalisms and methodologies, relationship between different approaches and applications. While a limit of half an hour per talk was initially established, we decided to leave the time for discussion relatively open, since several open topics and new developments were envisaged out of presentations.

Collective discussions have been arranged along four topics, i.e. basic concepts and foundations, specific formalisms for argumentation, algorithms, and connections both inside the argumentation field and with outside research topics.

We organized three discussion groups each headed by one organizer (see Section 4). Each group was asked to identify the most important open problems in argumentation. Interestly enough, there was little intersection between the three outcomes, i.e. the three groups came out with different problems. Many of them concerned foundational issues of the theory, e.g, how to formally represent various kinds of arguments and how to identify sets of postulates on the reasoning activity over arguments in specific contexts. On the other hand, the relationship between argumentation and other research fields (e.g. natural language processing, machine learning, human computer interaction, social choice) was seen to be of major importance, especially to develop more applications.

The unique setting and atmosphere of Dagstuhl provided the ideal environment to exchange ideas on future directions of argumentation, with discussions often lasting all the evening and the first part of the night.

The Perspectives Workshop concluded with the presentation of the results yielded by the group discussions, that in our opinion will lead to collaborative research, scientific papers and funded international projects in the future.

2 Table of Contents

3 Overview of Talks

3.1 Argumentation theory in formal and computational perspective

Bart Verheij (University of Groningen, NL)

Joint work of van Eemeren, Frans; Verheij, Bart
Main reference F. H.van Eemeren, B. Garssen, E. C. W. Krabbe, A. F. Snoeck Henkemans, B. Verheij, J. H. M. Wagemans, "Handbook of Argumentation Theory," ISBN 978-90-481-9472-8, Springer, 2014.
URL http://www.springer.com/de/book/9789048194728

As authors of a recent handbook of argumentation theory (not focused on the formal as is the present handbook), we have planned chapter 1 of the handbook of formal argumentation with three aims:

1. Introduce argumentation theory as an interdisciplinary research discipline.
2. Provide a bridge from informal to formal argumentation theory.
3. Aim at a readership of people with various backgrounds.

As such, the approach of the chapter tries to balance the kinds of methods, research styles and ideas, found across the triangle of cognitive systems:
bottom corner: Theoretical systems (philosophical paradigms, formalisms)
top-left corner: Artificial systems (software, robots)
top-right corner: Natural systems (texts, dialogs)

We hope the chapter can contribute to theoretical progress (growth towards standardized theory, connections with related theory) and applied progress (growth of relevant software support, collections of relevant knowledge). As a means, we suggest an enhanced exchange and collaboration between researchers of different backgrounds.

3.2 Historical Overview of Formal Argumentation

Henry Prakken (Utrecht University, NL)

The history of formal argumentation is described in terms of a main distinction between argumentation-based inference and argumentation-based dialogue. For both aspects of argumentation historical influences and trends are sketched.

3.3 Argumentation, nonmonotonic reasoning and logic

Alexander Bochman (Holon Institute of Technology, IL)

We provide s formal description of logical systems that can properly host various argumentation frameworks. It is shown, in particular, that the majority of such systems are representable as extensions of Dung's argumentation frameworks in suitable logical languages.

3.4 Abstraction1 vs. Abstraction2 in Formal Argumentation

Leendert van der Torre (University of Luxembourg, LU)

Joint work of Gabbay, Dov; Liao, Beishui; van der Torre, Leendert

We define abstraction1 as an equivalence relation over inputs to classify operators to compute conclusions, such as the use of dominance graphs in voting theory, and abstraction2 as a way of handling complexity, reusability, interoperability, and independence of implementation, such as the use of artificial languages in computer science, and (maybe) the use of natural language. Papers and theories about formal argumentation can be classified according to their stance towards abstraction. There are theories that do not consider abstract arguments, theories that consider both abstract1 and structured or instantiated arguments, and theories that consider abstract2 arguments only. We argue that research in these three classes is based on three distinct methodologies, and thus have distinct evaluation criteria. Though these two kinds of abstract1/2 argumentation theory are studied in two distinct volumes of the handbook on formal argumentation, we bring them here together in one chapter to look for common threads in the two disciplines, such as the role of refinement as the inverse of abstraction, and the role and use of auxiliary arguments. We consider also the role of fallacies in argumentation.

3.5 Requirements Analysis for Formal Argumentation

Tom Gordon (Fraunhofer FOKUS – Berlin, DE)

We suggest applying software engineering methods for "agile" requirements analysis to the development and evaluation of formal models of argumentation. The aim and purpose would be to help assure that formal models of argument are useful as a foundation for software tools supporting real argumentation tasks in domains such as law, politics and humanities scholarship and to help avoid developing a technical conception of "argument" far removed its meaning in fields of argumentation practice. We conclude with a list of some open issues and problems for which there are thus far no adequate formal models of argument, perhaps because prior research has not been sufficiently requirements driven.

3.6 Dung's traditional argumentation

Massimiliano Giacomin (University of Brescia, IT)

Joint work of Baroni, Pietro; Caminada, Martin; Giacomin, Massimiliano

This talk introduces Dung's argumentation frameworks and presents an overview on the semantics for abstract argumentation, including some of the most influential proposals. In particular, the talk reviews Dung's original notions of complete, grounded, preferred, and stable semantics, as well as subsequently proposed notions like semi-stable, ideal, eager,

naive, stage, CF2, stage2 and resolution-based semantics. Both extension-based and labelling-based definitions are considered. Furthermore, the talk reviews some general properties for semantics evaluation, analyzes the notions of argument justification and skepticism, and discusses the relationships among argumentation frameworks and their semantics. The final part of the presentation is focused on various lines of technical developments of Dung's model and open issues.

3.7 Abstract Dialectical Frameworks and Graph-Based Argument Processing

Gerhard Brewka (Universität Leipzig, DE)

Joint work of Brewka, Gerhard; Woltran, Stefan
Main reference G. Brewka, S. Woltran, "GRAPPA: A Semantical Framework for Graph-Based Argument Processing," in Proc. of the 21st Europ. Conf. on Artificial Intelligence (ECAI'14), Frontiers in Artificial Intelligence and Applications, Vol. 263, pp. 153–158, IOS Press, 2014.
URL http://dx.doi.org/10.3233/978-1-61499-419-0-153

Graphical models are widely used in argumentation to visualize relationships among propositions or arguments. The intuitive meaning of the links in the graphs is typically expressed using labels of various kinds. In this talk we introduce a general semantical framework for assigning a precise meaning to labelled argument graphs which makes them suitable for automatic evaluation. Our approach rests on the notion of explicit acceptance conditions, as first studied in Abstract Dialectical Frameworks (ADFs). The acceptance conditions used here are functions from multisets of labels to truth values. We define various Dung style semantics for argument graphs. We also introduce a pattern language for specifying acceptance functions. Moreover, we show how argument graphs can be compiled to ADFs, thus providing an automatic evaluation tool via existing ADF implementations. Finally, we also discuss complexity issues.

3.8 Abstract Rule-based Argumentation

Henry Prakken (Utrecht University, NL)

Joint work of Modgil, Sanjay; Prakken, Henry

First the standard ASPIC+ framework for structured argumentation is presented. Then several ways to use it are discussed, some variations of the framework are sketched and relations with other work are discussed.

3.9 Assumption-based argumentation

Pietro Baroni (University of Brescia, IT)

Joint work of Fan, Xiuyi; Schulz, Claudia; Toni, Francesca

The presentation describes the basic notions of ABA, its relationships with other formalisms, its syntax and semantics, the computational tool of dispute trees and the uses of the formalism for dialogues and explanation.

3.10 Argumentation Based on Logic Programming

Guillermo Simari (National University of the South – Bahía Blanca, AR)

Joint work of García, Alejandro J.; Simari, Guillermo R.

In this chapter, the connections between Logic Programming and Argumentation through the formalisms introduced in the literature are explored. These relations have enriched both areas contributing to their development. Some argumentation formalisms were used to define semantics for logic programming and also logic programming was used for providing an underlying representational language for non-abstract argumentation formalisms. Finally, different applications of the reasoning mechanisms based on argumentation in different areas of Artificial Intelligence such as Possibilistic Reasoning, Backing and Undercutting, Strength and Time, Decision Making, Planning, Ontologies, and Knowledge-based Systems are presented.

3.11 Constructing Argument Graphs with Deductive Arguments

Guillermo Simari (National University of the South – Bahía Blanca, AR)

Joint work of Besnard, Philippe; Hunter, Anthony

A deductive argument is a pair where the first item is a set of premises, the second item is a claim, and the premises entail the claim. This can be formalized by assuming a logical language for the premises and the claim, and logical entailment (or consequence relation) for showing that the claim follows from the premises. Examples of logics that can be used include classical logic, modal logic, description logic, temporal logic, and conditional logic. A counterargument for an argument A is an argument B where the claim of B contradicts the premises of A. Different choices of logic, and different choices for the precise definitions of argument and counterargument, give us a range of possibilities for formalizing deductive argumentation. Further options are available to us for choosing the arguments and counterarguments we put into an argument graph. If we are to construct an argument graph based on the arguments that can be constructed from a knowledgebase, then we can be exhaustive in including all arguments and counterarguments that can be constructed from the knowledgebase. But there are other options available to us. We consider some of the possibilities in this review.

3.12 Argumentation Schemes

Chris Reed (University of Dundee, GB)

Joint work of Macagno, F.; Reed, C.; Walton, D.

Argumentation schemes have been an influential component of both the philosophy and pedagogy of argumentation and critical thinking and also of formal and computational models of structured argumentation. In this chapter, we explore a number of issues relating to argumentation schemes. First, the challenges posed by critical questions are tackled, showing how different types of schemes correspond to different types of structure in both structured argumentation complexes and also in dialogical interactions. Next, we explore the connections between argumentation schemes and argument mining, including the particularly pernicious challenge of corpora and data management. As a part of this topic, the question of how nets of argumentation schemes can be composed. Finally, there is the issue of classification and organisation of schemes, whether taxonomically, ontologically, or on the basis of clusters, in order to provide clarity and structure for both practical and formal uses of argumentation schemes.

3.13 Rationality Postulates and Critical Examples

Martin Caminada (University of Aberdeen, GB)

We present the proposed structure of the chapter on Rationality Postulates and Critical Examples in the Handbook of Formal Argumentation.

1. Introduction
2. Preliminaries
3. Direct consistency, indirect consistency and closure
   a. restricted rebut solutions
      i. transposition
      ii. contraposition
      iii. semi-abstract approach of Dung and Tang
      iv. on the need of complete-based semantics
   b. unrestricted rebut solutions
4. Non-interference and crash resistance
   a. erasing inconsistent arguments
   b. requiring consistent entailment and forbidding strict-on-strict
5. Rationality postulates and other instantiations
6. Summary and discussion

3.14 Argument-Based Entailment as Discussion

Martin Caminada (University of Aberdeen, GB)

We describe the proposed structure of the chapter on Argument-Based Entailment as Discussion in the Handbook of Formal Argumentation.

1. Introduction
2. The preferred game
3. The stable game
4. The ideal game
5. The grounded games
   a. the standard grounded game (SGG)
   b. the grounded persuasion game (GPG)
   c. the grounded discussion game (GDG)
   d. overview and comparison
6. Discussion

3.15 On the Relation between AA, ABA and LP

Martin Caminada (University of Aberdeen, GB)

In the current talk, we examine the equivalences and differences between Assumption-Based Argumentation, Abstract Argumentation and Logic Programming. It is proposed that this could be the topic of an additional chapter in the Handbook of Formal Argumentation.

3.16 Computational Problems in Formal Argumentation and their Complexity

Wolfgang Dvorak (Universität Wien, AT)

Joint work of Dunne, Paul E.; Dvořák, Wolfgang

Several computational challenges arise in the process of formal argumentation. Understanding the computational complexity of these problems and different sources thereof is essential for the design of efficient argumentation systems that scale well with the size of argumentation scenarios. On a high-level there are three main tasks where computational challenges arise: (1) constructing arguments and identifying conflicts between them; (2) resolving the conflicts and identifying sets of coherent arguments; (3) drawing conclusions from the selected arguments. While the necessary computations in (1) and (3) are often purely in the underlying logic/formalism the tasks arising in (2) are argumentation problems at their core, and thus are often studied independently of a concrete instantiation of (1) and (3). We discuss three formalisms such that the different computational aspects are covered, namely Dung's Abstract Argumentation Frameworks, Assumption-based Argumentation (ABA)

and Abstract Dialectical Frameworks (ADFs). The complexity of reasoning tasks highly depends on the applied semantics and we categorize semantics by different levels of complexity (by their location in the so-called polynomial hierarchy) which is in accordance with the performance of existing argumentation systems for different semantics. As most of these problems are of high worst-case complexity, we also consider properties of instances, like being in a specific graph class, that reduce the complexity. Finally, we also show techniques from parametrized complexity that allow for a more fine-grained complexity classification taking structural properties into account.

References

1 Yannis Dimopoulos, Bernhard Nebel, and Francesca Toni. On the computational complexity of assumption-based argumentation for default reasoning. *Artif. Intell.*, 141(1/2):57–78, 2002. DOI: http://dx.doi.org/10.1016/S0004-3702(02)00245-X.

2 Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

3 Paul E. Dunne. Computational properties of argument systems satisfying graph-theoretic constraints. *Artif. Intell.*, 171(10-15):701–729, 2007.

4 Paul E. Dunne and Michael Wooldridge. Complexity of abstract argumentation. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 85–104. Springer US, 2009. ISBN 978-0-387-98197-0. DOI: http://dx.doi.org/10.1007/978-0-387-98197-0_5.

5 Wolfgang Dvořák. *Computational Aspects of Abstract Argumentation*. PhD thesis, Vienna University of Technology, Institute of Information Systems, 2012. DOI: http://permalink.obvsg.at/AC07812708.

6 Hannes Strass and Johannes Peter Wallner. Analyzing the computational complexity of abstract dialectical frameworks via approximation fixpoint theory. *Artif. Intell.*, 226:34–74, 2015. DOI: http://dx.doi.org/10.1016/j.artint.2015.05.003.

7 Johannes P. Wallner. *Complexity Results and Algorithms for Argumentation – Dung's Frameworks and Beyond*. PhD thesis, Vienna University of Technology, Institute of Information Systems, 2014. DOI: http://permalink.obvsg.at/AC11706119.

3.17 Implementations

Matthias Thimm (Universität Koblenz-Landau, DE)

We survey both the current state-of-the-art of general techniques and specific software systems for solving tasks in abstract argumentation frameworks, structured argumentation frameworks, and approaches for visualizing and analysing argumentation. Furthermore, we discuss general challenges and further promising techniques for solving these problems such as parallel processing and approximation techniques. Finally, we address the issue of evaluating software systems empirically with linkage to the International Competition on Computational Models of Argumentation.

3.18 Advanced techniques

Ringo Baumann (Universität Leipzig, DE)

The aim of the talk is to give an overview of fundamental properties of abstract argumentation frameworks typically considered for nonmonotonic formalisms. In particular, we shed light on the following issues/questions:

1. *Replaceability:* Is it, and if so how, possible to *simplify* parts of a given AF F, s.t. the modified version F' and F cannot be semantically distinguished by further information which might be added later to both simultaneously?
2. *Expressibility:* Is it, and if so how, possible to *realize* a given candidate set of extensions within a single AF F?
3. *Existence and uniqueness:* Is it, and if so how, possible to decide (without computing) whether a certain AF *possesses* an acceptable set of arguments w.r.t. a certain semantics? Moreover, in what situation the solution is *unique*? We study these questions for three classes of AFs, namely finite, finitary as well as the unrestricted case of arbitrary AFs.

3.19 A principle based evaluation of argumentation semantics

Leendert van der Torre (University of Luxembourg, LU)

Joint work of van der Torre, L.; Vesic, S.

This chapter gives a classification of argumentation semantics based on a set of principles. Starting from Baroni and Giacomin's original classification, we extend their analysis with other semantics and principles proposed in the literature.

3.20 Locality and Modularity in Abstract Argumentation

Pietro Baroni (University of Brescia, IT)

Joint work of Baroni, Pietro; Giacomin, Massimiliano; Liao, Beishui

The presentation discusses the motivations for investigating locality and modularity properties in abstract argumentation and surveys the main results available in the literature concerning directionality, SCC-recursiveness and decomposability and their uses for efficient computation, interchangeability and summarization.

3.21 Semantic Instantiations

Emil Weydert (University of Luxembourg, LU)

Formal argumentation is characterized by diverging accounts and a number of controversial issues. This raises the question of validation and common foundations for an area which in the past had a mainly proof-theoretical flavour. The present chapter discusses approaches trying to semantically ground argument systems and argumentation-based reasoning. In fact, arguments can be interpreted as constraints over epistemic states. Adopting a very general perspective where arguments are seen as inferential graphs over a defeasible conditional logic, it becomes possible to exploit powerful semantic techniques from default reasoning. A proptotypical instance are Dung-style acceptance functions based on the ranking measure semantics for default inference.

3.22 Processing Argumentation in Natural Language Texts

Katarzyna Budzynska (Polish National Academy of Sciences, PL, and University of Dundee, UK)

Joint work of Budzynska, Katarzyna; Villata, Serena

Discourse analysis and text mining is a promising approach to identify and extract real-life arguments, receiving attention from the natural language processing community (e.g., argument mining of legal documents, on-line debates, newspaper and scientific articles, etc). On the other hand, computational models of argumentation have made substantial progress in providing abstract and structured formal models to represent and reason over argumentation structures. Our work is aimed at the interaction between Computational Linguistics and Argumentation Theory. More precisely, it has the goal to combine the techniques and frameworks for analysing, aggregating, synthesizing, structuring, summarizing, and reasoning about arguments in natural language texts.

4 Working Groups

4.1 Results of Discussion Group I – Most Important Problems in Argumentation

Beishui Liao (Zhejiang University, CN)

- The role of numerical approaches
- Interaction and aggregation of arguments
- Formal representation of argument
- How to use argumentation to represent preference-based nonmonotonic reasoning
- What is the negation of argument

- Formal argumentation account of fallacies
- Analysing and modelling argumentation schemes
- Argumentation models of decision theory vs. other models of decision making
- Argumentation mining (e.g. large-scale applications)
- Argumentation and other networks
- Validity of arguments w.r.t. time/dynamics
- Balancing the different steps of argumentation

4.2 Results of Discussion Group II – Most Important Problems in Argumentation

Massimiliano Giacomin (University of Brescia, IT)

- How to do reasoning with strict and defeasible (non-strict) rules by satisfying qualitative postulates and in a way which is expressible dialectically in a natural way, without being overly skeptical?
- Identifying proper sets of qualitative postulates that should be satisfied in specific contexts.
- Alternatives to Dung's approach. Identifying an elegant formalism encompassing Dung's model and capturing also different ways of evaluating arguments, e.g. balancing considerations.
- Achieving a clarification on the "semantics of a semantics". When to adopt a specific semantics instead of another?
- How do we validate dialogue protocols? Do we need a semantic model?
- How to do sound and complete argument games when arguments become available dynamically from private knowledge bases?
- Identifying models to switch between different levels of reasoning, as happens in real argumentation.
- What is the nature of defeat? How to deal with preferences? Preference order between arguments is dynamic and may depend on the labelling of arguments, thus a recursive process may be needed.
- Further investigation on the notion of accrual and its management.
- How to manage numerical information in argumentation in a principled way?
- How to determine "who knows more" in a multi-agent argumentation context? What is knowledge? Relationships with other areas (e.g. belief revision and logic programming).
- Dealing with time in argumentation, e.g. arguments can be valid now but not in the future.

4.3 Results of Discussion Group III – Most Important Problems in Argumentation

Leendert van der Torre (University of Luxembourg, LU)

- What is the negation of an argument? Type theory. Define the operators, like negation of trust in distrust, negation of attack as support, negation of argument. What is the negation of an argumentation framework?
- Interaction of strict and non-strict rules in argumentation. Do we need strict rules? What are strict rules? How does it relate to work in general NMR? What is the role of specificity in this discussion? Relationship with the work by Frida Stolzenburg, last year at KR, David Poole's approach.
- Relationship between argumentation and natural language processing / machine learning / data mining. Output of many techniques are different from an argumentation framework. For example, argument mining, learning strategies, giving reasons for what it learned. Define more clearly the argument mining problem, extend the interdisciplinary
- Rhetorics and dialectics, debating game to beat politician
- Natural language interfaces to arguments.
- Integrating argumentation and computational social choice. The relation between voting and the semantics of argumentation. Show that semantics works better. Kind of democracy based on argumentation.
- Can argumentation contribute to Turing test, Winograd scheme, disambiguating sentences, giving reasons why one way or another . Is AI a sub field of machine learning? Relationship with the other Dagstuhl workshop. How do we convince Russell and Norvig that formal argumentation should be in the book?
- Alternatives to the three step approach, sometimes we are interested in only one argument, focus on explanation and justification
- Bringing argumentation to the U.S. (Kevin Ashley, Thorne MacCarthy)
- What is rationality and which is the role formal argumentation
- Efficient algorithms for abstract argumentation not based on SAT problem

Participants

- Pietro Baroni
University of Brescia, IT
- Ringo Baumann
Universität Leipzig, DE
- Stefano Bistarelli
University of Perugia, IT
- Alexander Bochman
Holon Institute of Technology, IL
- Gerhard Brewka
Universität Leipzig, DE
- Katarzyna Budzynska
Polish Academy of Sciences – Warsaw, PL
- Martin Caminada
University of Aberdeen, GB
- Federico Cerutti
University of Aberdeen, GB
- Wolfgang Dvorak
Universität Wien, AT
- Dov M. Gabbay
King's College London, GB
- Massimiliano Giacomin
University of Brescia, IT
- Tom Gordon
Fraunhofer FOKUS – Berlin, DE
- Beishui Liao
Zhejiang University, CN
- Henry Prakken
Utrecht University, NL
- Chris Reed
University of Dundee, GB
- Odinaldo Rodrigues
King's College – London, GB
- Guillermo R. Simari
National University of the South – Bahía Blanca, AR
- Matthias Thimm
Universität Koblenz-Landau, DE
- Leendert van der Torre
University of Luxembourg, LU
- Bart Verheij
University of Groningen, NL
- Emil Weydert
University of Luxembourg, LU
- Stefan Woltran
TU Wien, AT