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