Report from Dagstuhl Seminar 17491

Computational Metabolomics: Identification, Interpretation, Imaging

Edited by

Theodore Alexandrov[1], Sebastian Böcker[2], Pieter Dorrestein[3], and Emma Schymanski[4]

1 **EMBL Heidelberg, DE,** theodore.alexandrov@embl.de
2 **Universität Jena, DE,** sebastian.boecker@uni-jena.de
3 **UC – San Diego, US,** pdorrestein@ucsd.edu
4 **University of Luxembourg, LU,** emma.schymanski@uni.lu

Abstract

Metabolites are key players in almost all biological processes, and play various functional roles providing energy, building blocks, signaling, communication, and defense. Metabolites serve as clinical biomarkers for detecting medical conditions such as cancer; small molecule drugs account for 90 % of prescribed therapeutics. Complete understanding of biological systems requires detecting and interpreting the metabolome in time and space. 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 metabolite mixtures. MS-based experiments generate datasets of increasing complexity and size.

The Dagstuhl Seminar on Computational Metabolomics brought together leading experts from the experimental (analytical chemistry and biology) and the computational (computer science and bioinformatics) side, to foster the exchange of expertise needed to advance computational metabolomics. The focus was on a dynamic schedule with overview talks followed by break-out sessions, selected by the participants, covering the whole experimental-computational continuum in mass spectrometry. Particular focus in this seminar was given to imaging mass spectrometry techniques that integrate a spacial component into the analysis, ranging in scale from single cells to organs and organisms.

Seminar December 3–8, 2017 – http://www.dagstuhl.de/17491
1998 ACM Subject Classification J.3 Life and Medical Sciences
Keywords and phrases algorithms, bioinformatics, cheminformatics, computational mass spectrometry, computational metabolomics, databases, imaging mass spectrometry
Digital Object Identifier 10.4230/DagRep.7.12.1
Edited in cooperation with Marcus Ludwig

1 Executive summary

Theodore Alexandrov
Sebastian Böcker
Pieter Dorrestein
Emma Schymanski

Metabolomics is the study of metabolites (the small molecules involved in metabolism) in living cells, cell populations, organisms or communities. Metabolites are key players in

Computational Metabolomics: Identification, Interpretation, Imaging, *Dagstuhl Reports*, Vol. 7, Issue 12, pp. 1–18
Editors: Theodore Alexandrov, Sebastian Böcker, Pieter Dorrestein, and Emma Schymanski
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

almost all biological processes, play various functional roles providing energy, building blocks, signaling, communication, and defense and serve as clinical biomarkers for detecting medical conditions such as cancer. Small molecule drugs (many of which are derived from metabolites) account for 90 % of prescribed therapeutics. Complete understanding of biological systems requires detecting and interpreting the metabolome in time and space.

Mass spectrometry is the predominant analytical technique for detecting and identifying metabolites and other small molecules in high–throughput experiments. Huge technological advances in mass spectrometry and experimental workflows during the last decade enabled novel investigations of biological systems on the metabolite level. Research into computational workflows, the simulation of tandem mass spectra, compound identification and molecular networking have helped disentangle the vast amount of information that mass spectrometry provides. Spatial metabolomics on different spatial scales from single cells to organs and organisms has posed data analysis challenges, in particular due to an unprecedented data volume generated that grows quadratically with the increase of spatial resolution.

Continued improvements to instruments, resolution, ionization and acquisition techniques mean that metabolomics mass spectrometry experiments can generate massive amounts of data, and the field is evolving into a "big data" science. This is particularly the case for imaging mass spectrometry, where a single dataset can easily be many gigabytes or even terabytes in size. Despite this dramatic increase in data, much of the data analysis in metabolomics is still performed manually and requires expert knowledge as well as the collation of data from a plethora of sources. Novel computational methods are required to exploit spectral and, in the case of imaging, also spatial information from the data, while remaining efficient enough to process tens to hundreds of gigabytes of data.

Dagstuhl Seminar 17491 on Computational Metabolomics: Identification, Interpretation, Imaging built on the success of the first Computational Metabolomics Dagstuhl Seminar (15492) in 2015. A number of topics overlapped with the 2015 seminar, while the focus on imaging introduced new perspectives, participants and topics. In contrast to the first seminar, 17491 was a large seminar, with 45 very active participants and a large portion of young scientists. From the first hours of the seminar, effort was made to integrate these young scientists in the discussions and presentations and this paid off leading to lively discussions involving all participants. Many participants were new to Dagstuhl and the concept of Dagstuhl seminars, which led to a seminar that was a combination of being semi-structured and spontaneous. Very positive feedback was received from all during a comprehensive feedback session before lunch on Friday, including constructive ideas for a new focus for a possible new seminar in 2019.

On the scientific side, the seminar covered numerous topics which were found to be most relevant for the computational analysis of mass spectrometry data, and ranged from the "dark matter in metabolomics" to "integrating spatial and conventional metabolomics"; see the full report for a comprehensive description.

The seminar has fully achieved its key goals: to foster the exchange of ideas between the experimental and computational communities; to expose the novel computational developments and challenges; and, to establish collaborations to address grand and priority challenges by bridging the best available data with the best methods.

2 Table of Contents

3 Overview of Talks

3.1 Challenges in "conventional" metabolomics

Corey Broeckling (Colorado State University – Fort Collins, US), and Nicola Zamboni (ETH Zürich, CH)

Metabolomics is a field of research which relies on a broad collection of preparation and analytical approaches and techniques. We began this conference by outlining the range of approaches used and provide a quick overview of the problems with metabolomics that may be addressed using computational approaches. We identified computational opportunities in the following areas:

1. Increase in breadth and coverage of metabolomics
2. workflows for efficient, reusable and objective annotation
3. processing standards for interoperability and testing
4. systematic analysis of MS features
5. network-driven data mining
6. standardization / normalization of non-targeted metabolomics

3.2 Challenges in spatial metabolomics

Theodore Alexandrov (EMBL Heidelberg, DE)

Spatial metabolomics is about capturing metabolome in its full complexity across various spatial scales. Currently there are several methods, in particular imaging mass spectrometry, which generate 1 TB of data per sample. Challenges are:

1. the interpretation of data,
2. data curation, and
3. multi-omics integration.

This requires cloud computing and collaboration.

3.3 Challenges in environmental exposomics and environmental cheminformatics

Lee Ferguson (Duke University – Durham, US)

Identification of unknown pollutants and toxicants in environmental and biological samples is complicated by incomplete molecular databases. Difficulty in prioritizing chemicals used in commerce or causing adverse effects. Advances in mass accuracy, resolution and data acquisition rates have increased data acquisition rates, but annotation remains difficult. Priorities for future advancements include the incorporation of false discovery rates for

both *in silico* and library-based MS/MS identification strategies. Incorporation of metadata for molecular candidate prioritization will be critical to enhancing identification rates in environmental samples. Computational methods for establishing networks associated with molecules in environmental systems will be vital for understanding relationships among pollutants.

Emma Schymanski (LCSB, University of Luxembourg, LU)

The dark matter in environmental (and small molecule) samples is still a huge challenge. We have untreated wastewater discharged into rivers with measurable toxic effects that can not be clarified with known (target) chemicals. Homologues (UVCBs, complex mixtures) are a massive part of this dark matter with thousands of homologous series. On the other hand we have 100,000s of complex chemical mixtures produced in thousands of tonnes where we cannot even assign a structure to the complex name. How can we reconcile this?

3.4 False discovery rate estimation

Sebastian Böcker (Universität Jena, DE)

FDR allows for an automated, objective and reproducible estimation of "how unsure we are": This is accepting the fact that in science, there is no "absolutely sure". It is build on the assumption that the score of the hit is allowing us to discriminate between true hits (correct identifications) and bogus hits (incorrect identifications). For metabolomics we face a number of issues and challenges:

1. ID rates are still much smaller for *in silico* tools than in, say, proteomics.
2. Spectral libraries are notoriously incomplete.
3. Separation by score is significantly worse than, say, in proteomics.
4. We have no ideas how to generate decoy structures, or how to transform them into fragmentation spectra.

Andrew Palmer (EMBL – Heidelberg, DE)

Many scoring systems exist for measuring the quality of match between experimental data and reference databases, for example tandem spectra of isotope patterns. At some point a threshold must be established for those scores in order to separate the comparisons into "interesting" and "uninteresting". We discussed some newly developed approaches for estimating the false discovery rate for such comparisons in metabolomics experiments, in particular the prediction and evaluation of target-decoy approaches and the requirements for accurate estimation. It is clear that no approach is perfect but the community agrees that quantifying database matching performance is essential.

3.5 Computational challenges in environmental metabolomics and exposomics

David Wishart (University of Alberta – Edmonton, CA)

Humans are exposed to all kinds of chemicals throughout their lifetimes. These "environmental" exposures account for many of the chronic diseases that develop later in life. In the USA, more than 90% of all deaths (in 2013) could be attributed to some kind of chemical, biological or environmental exposure. The measurement of chemical exposures – both within the body and outside the body – is called exposomics. In this presentation I presented a brief overview of exposomics and identified 6 key challenges that are facing the field. These include:

1. the problem with automated workflows
2. the missing "pure" compound problem
3. the missing "metabolized" compound problem
4. the missing "observable" problem
5. the missing ontology problem and
6. the missing funding problem.

Potential solutions for each of these challenges are presented.

3.6 Improved molecular networks with LC-MS feature detection and *in silico* annotation

Louis-Felix Nothias-Scaglia (UC – San Diego, US)

Recent advances and challenges in MS-data preprocessing for molecular networking on GNPS web-platform (http://gnps.ucsd.edu) were presented and discussed during the session. This processing step improves qualitatively the molecular networks by filtering-out noisy features, by reducing the data redundancy, and by enabling the discrimination of isomeric ions based on the retention time. Additionally, this approach allows the integration of semi-quantitation in the network which is a key for new mining strategy, such as the "bioactive molecular networking" (Nothias, J. Nat. Prod., 2018, accepted). One of the most exciting outcome of that development relies, is the possibility of using *in silico* annotation tools such as Sirius / CSI:FingerID (Dührkop, PNAS, 2015) or Network Annotation Propagation. The biggest challenge remains the need to optimize LC-MS feature detection parameters on a dataset-basis, which hamper the systematic large-scale use of that approach on all public dataset. The development of a LC-MS processing tool that would include an "auto-tuning" feature is needed to solve that bottleneck.

3.7 Issues in MS1 data processing and annotation

Julijana Ivanisevic (University of Lausanne, CH), Michael Andrej Stravs (Eawag – Dübendorf, CH)

A presentation and subsequent discussions highlighted current unsolved issues in MS1 data processing, including:

- feature detection – i.e. chemical and bioinformatics noise
- feature annotation – issues with componentization, i.e. grouping of isotopes, adducts etc.
- batch correction as a problem of both experimental and computational approaches.

In the presentation it was pointed out that only small portion of acquired MS1 data is indeed assigned as isotopes, adducts, etc. The amount of non-annotated MS1 data remains extremely high (at least – more than a half of detected signals) implying the presence of noisy features (i.e. detector artifacts, data artifacts) but especially that the redundancy is still poorly annotated like in the case of multiple charged species, in-source fragments, etc. Discussions highlighted the need for solid annotated test data necessary for the development of improved computational approaches for feature detection and annotation.

3.8 *In silico* structure prediction with CSI:FingerID

Kai Dührkop (Universität Jena, DE)

Identifying metabolite structures via tandem MS is one of the main challenges in metabolomics. *In silico* / combinatorial fragmenters start from a hypothetical structure (taken from a structure database) and match it against the measured spectrum, reporting some kind of likelihood or match score. In contrast, CSI:FingerID predicts a hypothetical structure (in form of a probabilistic molecular fingerprint) directly from the measured spectrum using machine learning techniques. This allows to deal with so called "unknown unknowns" - structures which are not contained in any structure database. But it is also a starting point to extract knowledge from MS/MS data without the necessity to identify the exact structure. We discussed about visualization of predicted structures and possible applications for structure prediction.

3.9 Long-term monitoring of aquatic systems, combining LC-HRMS and chemometric tools to highlight organic pollutants

Martin Loos (looscomputing – Dübendorf, CH)

Liquid chromatography coupled to high-resolution mass spectrometry has become the method of choice to trace highly diluted (yet possibly toxic) organic micropollutants. Despite their widespread release, data mining workflows to prioritize trends of concern with respect to

increasing pollutant concentrations have remained scarce within an environmental monitoring context. Here, we have presented certain steps within any such workflow to automatize, streamline and facilitate the fast detection of such patterns, even from n $\geq$ 1000 LC-HRMS files.

3.10 Topic modelling for substructure discovery in metabolomics data: state of the art and challenges ahead

Justin van der Hooft (Wageningen University, NL)

Mass Spectrometry-based metabolomics workflows result in large amounts of data often containing fragmentation spectra of many detected molecules. Ever since fragmentation spectra of biomolecules could be produced, data analysts have been looking for specific fragmentation patterns that they could couple to key structures or substructures in their samples. Doing so, they could start to structurally annotate the molecules in a complex biological mixture. However, the manual analysis of MS/MS spectra is tedious and impossible when faced with over 5000 MS/MS spectra for each sample. To overcome this hurdle, computational approaches have been proposed over the last years. The application of topic modelling, originally used for text-mining, to MS/MS data was recently introduced. In this talk, the MS2LDA algorithm was introduced: concurring mass fragments and/or neutral losses defined from fragmented molecules are discovered and grouped into Mass2Motifs (similar to topics in text-mining). This is the first unsupervised approach that enables the detection of potential substructures in mass spectrometry fragmentation data. Validation results using standards from MassBank and GNPS were shown, as well as Mass2Motifs discovered in beer samples. Indeed, MS2LDA found biochemically relevant substructures that could be annotated with amino acid, sugar, and aromatic moieties, amongst others. Furthermore, it was shown that MS1 comparisons can be mapped on the Mass2Motifs, thereby guiding the user to relevant/interesting Mass2Motifs, for example, those more present or absent in Indian Pale Ale (IPA) beers. Finally, it was discussed how to best take this approach forward, in particular regarding annotation of the discovered fragmentation patterns. As is true for text-mining, the discovered Mass2Motifs (topics) are a collection of fragments/losses (words) that need to be interpreted by the user. It was concluded that further integration with other tools and efficient storage of annotated Mass2Motifs are prerequisite for full exploitation of this innovative approach.

3.11 Integrating mass spectrometry with other imaging modalities: Improving biological insight through data-driven multi-modal image fusion

Raf Van de Plas (TU Delft, NL)

Studies in medicine and biology increasingly employ a multitude of different imaging technologies to answer a specific biological question. A growing number of such multimodal imaging studies include imaging mass spectrometry (IMS) as one of these modalities. Although different modalities are routinely registered and overlaid to generate a single display, true integration of data across technologies is largely left to human interpretation, resulting in a significant underutilization of the potential of multi-modal measurements. This talk gives an overview of our recent work on the integration or "fusion" of IMS with measurements from other imaging modalities (Van de Plas et al., Nature Methods, 2015) and demonstrates the potential of data driven image fusion for IMS through several predictive applications. Example applications include:

1. the "sharpening" of IMS images, using microscopy measurements to predict ion distributions a la spatial resolution that exceeds that of measured ion images by ten times or more;
2. the enrichment of biological signals and the removal of instrumental noise by multi-modal corroboration; and
3. the prediction of ion distributions in tissue areas that were not-measured by IMS.

We also highlight more recent work in which contrary to fusing IMS with microscopy, our data-driven fusion method is used to combine liver mass spectrometry-based modalities into a single predicted modality that combines advantages of the several modalities. In this new IMS-IMS fusion setting, MALDI-FTICR IMS measurements (lower spatial resolution, higher mass resolution) enabling ion distributions to be predicted with both high spatial as well as high mass resolution. Examples are shown in lipid imaging, where there is both a need to spatially resolve fine tissue structure, as well as a need for high chemical specificity due to nominal isobaric species.

4 Working groups

4.1 Metabolite structure ambiguity – representation, standardization, and naming

Nils Hoffmann (ISAS – Dortmund, DE)

The structure of small molecules can not be fully established with current MS/MS methods. Ideally, structures should be unambiguously resolvable down to the isomer level, however, current mass spectrometric methods are limited to measuring accurate masses of precursor and fragment ions. MSn fragmentation patterns can assist in the elucidation of structures, but still fail to identify e.g. the positions of double bonds or specific ligands, since the corresponding fragment masses are virtually identical.

For lipids specifically, ambiguities exist concerning the position of double bonds in lipid chains and how to encode them in a consistent naming scheme. The current nomenclature uses lipid category abbreviations (e.g. PC) and encodings for the number of Carbon atoms in the fatty acid (FA) side chains (R1, R2, R2 ...) of a lipid and optionally, the number of double bonds (unsaturated bonds) in them, e.g. for a triglycerol with a total of 52 carbon atoms and one double bond, the name would be reported as TG(52:1), or as TG(16:0_18:0_18:1) if is known that one of the fatty acid chain consists of sixteen carbon atoms, the second FA chain of eighteen, and the third one of eighteen carbon atoms with one double-bond at an arbitrary position.

In principle, the same issues arise not only in lipidomics, but similarly for other "small molecules" (<1kDa) in environmental chemistry, glycomics, natural product chemistry (flavonoids and terpenes) and metabolomics. We identified the following, cross-cutting requirements for a nomenclature and representation of incomplete or ambiguous structural information:

1. representation of such information as extended SMILES / SMARTS and InChI (with extension of the current standard),
2. visualization of generalised structures, e.g. using CDK-depict, based on extended SMILES or other structural representations,
3. curation and availability of uncertain structures in databases,
4. support for reporting of ambiguity / structural uncertainty in community data standards, e.g. in mzTab, mzIdentML etc.,
5. search of patterns, e.g. conserved and variable substructures in structure databases
6. collapsing / superposition of defined, unique structures represented as a common conserved and a "common" varying part (e.g. exact ligand positions in Markush structures).

We want to address these requirements by gathering examples from the different communities that reflect the status quo of reporting ambiguous, not fully-resolved structures, especially when identification is only based on MS1 / MS2 database identification. We intend to use those to illustrate the benefit of having a common standard in a statement article that defines the problems and shortcomings of current reporting of structural information and raises awareness in and receives input from the affected communities.

We will work towards better interoperability between the tools used to generate extended SMILES (support for Rs, *s, etc. in the CDK, ChemAxon, and OpenBabel) and to support the encoding of uncertainty in ligand positions, e.g. for aromatic compounds, and to visualize them.

4.2 R-based computational mass spectrometry

Michael Andrej Stravs (Eawag – Dübendorf, CH)

Features and benefits of the base package MSnlib for treatment of mass spectrometry data were introduced, with a particular focus on the classes for representation of MS1 and MS2 spectra. A brief overview about features and interface of the new XCMS3 version were presented. A discussion highlighted a current gap in packages for MS2 library management and search, and adoption/extension of current approaches was discussed. Finally, the current trend of "tidyverse" packages was briefly exposed, and benefits and drawbacks of a possible adoption of "tidyverse" data structures were discussed.

4.3 Connecting genome and metabolome data: combining structural information from genome and metabolome mining

Justin van der Hooft (Wageningen University, NL)

Metabolomics workflows result in the discovery of molecular families sharing common core structures. Genome mining workflows result in the prediction of biosynthesis gene clusters responsible for the production of specialized molecules – those gene clusters can then be grouped in gene cluster families that produce similar molecules. Connecting genome and metabolome mining workflows can enhance the structural and functional annotation and identification of both metabolites and genes. By linking existing and novel networking approaches, accelerated substructure annotation can be accomplished, which will facilitate structural elucidation of specialized molecules. Through the link with the genome, the microbial producers of these molecules can also be linked. This break-out session discussed ideas and major challenges to exploit the linkage between these two largely separately developed omics fields: the availability of paired data sets including validated links between genome and metabolome; the type of statistics needed to correct for potential bias when correlating presence/absence of molecular and gene cluster families across different strains; the application of fragmentation trees to find substructures connecting to genes as well as information on how substructures are linked; the application of biotransformation rules to both substructures and complete structures; the use of (a subset of) CSI:FingerID substructures to assess the likelihood that they are present in fragmented molecules; and the possibility to also use transcriptomics data to assess if BGC are active or silent. It was concluded that the currently available paired data sets are scattered and not easily accessible: cross-linking tables and validated links are needed to apply machine learning tools. Furthermore, both genome and metabolome mining are highly evolving fields that could benefit from integration; examples for peptide-based specialized molecules were mentioned; however, for most classes novel tools are needed. Exciting ideas to apply machine learning to combine specific structural information from genome and metabolome mining were also shared.

4.4 Dark Matter

Ricardo Da Silva (UC – San Diego, US)

Dark matter was defined as frequently observed spectral signals for which no annotation can be assigned. The first source of dark matter was attributed to experimental design, highlighting the importance of having controls, in order to differentiate real signal from noise and contamination. The second source was attributed to spectrometric data pre-processing, due to signal complexity (adducts, isotopes, fragments, contaminants) and also due to the low accuracy of existing tools (missing peaks, integration of noise, split peaks, merged peaks). The third source was attributed to the limited coverage of spectral libraries, data repositories, specially for benchmark datasets, and the lack of connection between the known chemicals and its biological source. The most important long term solutions cited were the expansion of reference and raw data databases, as well as the design and analysis of benchmark datasets by multiple orthogonal methods and multiple labs.

4.5 Retention time and retention time prediction

Michael Witting (Helmholtz Zentrum – München, DE)

Retention time (RT) represents an orthogonal information to mass spectrometry. It is especially important to distinguish isomers which cannot be separated solely by MS and MS/MS. In this beak-out group several small presentations showed the state-of-the art in reporting RT data and predicting it. Different ways of improving RT prediction were discussed and important parameters to model retention time order were collected. These will serve as future reference for collecting metadata and data around RT and its prediction.

4.6 Data independent acquisition

Corey Broeckling (Colorado State University – Fort Collins, US)

Several members discussed the increasingly common 'data independent acquisition' (DIA) of MS/MS data. A workflow was presented for processing 'all ion fragmentation' data, an acquisition approach in which no precursor selection is performed. The various flavors of data independent acquisition methods that have been published were summarized graphically, in an effort to inform developers of DIA processing workflows to design processing tools which are sufficiently versatile to handle the many iterations. A considerable effort is being made to update and expand the ramclustR package with novel clustering methods, and a novel R based workflow was presented demonstrating efficacy in using extracted ion chromatograms to remove contaminating signals from DIA spectra. More broadly, the attendees discussed the many strengths of DIA approaches, as well as the frequent tradeoff between sensitivity and selectivity, and the difference between actual (physical isolation defined by precursor isolation window size) and processing assisted selectivity (overlapping windows can enable contaminant signal subtraction).

4.7 Molecular networking and integration with annotation tools

Madeleine Ernst (UC – San Diego, US)

Mass spectral molecular networks enable the observation of similarities and differences in MS/MS fragmentation patterns of complex samples. MS/MS fragmentation patterns are typical of a molecular structure, and molecular structures can thus be identified/annotated manually or by using *in silico* approaches. This break-out session discussed ideas and challenges to integrate molecular networks with *in silico* annotation tools. Mass spectral molecular networking has been made widely accessible by Global Natural Products Social Molecular Networking (GNPS), allowing the community to share, analyze and annotate MS/MS data. Nevertheless, mass spectral molecular networking is currently a per-study

approach and comparison of different datasets is only possible with limitations. Major challenges discussed in this session ranged from questions on how to integrate information on mass shifts in an automated way, simulate enzymatic reactions to predict molecular structures and incorporate biochemical properties (e.g., pH, bioactivity) in the networks to using network topology to improve *in silico* annotation. The need for sharing data with the community and making datasets public was stressed as well as the aim of integrating all types of information into the mass spectral molecular networks, ultimately enabling inter-study comparisons.

4.8 Bridging the gap

Sarah Scharfenberg (IPB – Halle, DE)

Metabolomics science faces several gaps, such as language and communication problems between the tool developers and the experimentalist, insufficient visibility of available tools and inappropriate usability of tools. Collaborative projects will fail as soon as one of the involved parties assesses its part as a service. Early communication of study design should be mandatory for each experimentalist who wants statistical support. To match the needs of both sides, we should further go for user driven development, which is based on a real study and a fixed problem and accompanied and constantly feedbacked by the corresponding experimentalist. To improve the usage of existing tools we could provide more educational documentation, such as webinars, video tutorials and example datasets.

4.9 Creating the "perfect" benchmark / reference dataset

Corey Broeckling (Colorado State University – Fort Collins, US)

The "perfect benchmark dataset" discussion began by trying to define the computational problems in data processing, including being able distinguish signal from noise and metabolites from artifacts. The concept of theoretical data and what sort of artifacts we might be able to predict/model was discussed, in an effort to determine how realistic the predicted data would be. A small "ring trial" was proposed, with a sample set based on the notion of a sample set containing a moderately complex (20 - 100) set of pure authentic standard compounds. This compound mixture will be analyzed under realistic conditions as if it were an authentic sample set. The sample mixture could additionally be sent to interested labs for analysis. Computational collaborators would be delivered data from one or more laboratories rather than samples. Participants intend to move the concept forward with a more detailed planning document.

4.10 Integrating spatial and conventional metabolomics

Andrew Palmer (EMBL – Heidelberg, DE)

The computational communities that focus on traditional metabolomics and imaging mass spectrometry have both approached the large scale processing of mass spectra from different perspectives. This session was an opportunity for both communities to describe their experiments, data, and processing strategies to seek opportunities for the exchange of ideas and methodologies. After all participants had an opportunity to present their methodologies, discussion focussed on the similarities between imaging mass spectrometry and high throughput flow injection experiments: both in terms of the experimental aims (to maximally annotate the peaks present) and data (large numbers of high resolution mass spectra from complex mixtures). To begin to assess the efficacy of computational strategies across these modalities a source of well characterised publically available data from flow injection studies needs to be identified.

4.11 *Ab initio* network reconstruction challenges

Fabien Jourdan (INRA-ENVT – Toulouse, FR)

First part of the discussion was about sharing views on network research field related to metabolomics. In fact, "network" is a very versatile concept used for many applications from molecular networks (nodes represent fragmentation spectra and edges cosine score) to biochemical networks (nodes are compounds and edges correspond to metabolic reactions). It is thus of utmost importance to define carefully what nodes and edges are modelling. Discussion then focused on the lack of identifier standardisation between modelling and metabolomics community. In particular, as a community, we advice that more care should not be taken when curated metabolic networks in order to provide InChIKeys and InChIs. The other issue, related to other break-out session, is the necessity of providing identifiers with flexible substructures identifiers (e.g. to deal with class of lipids). Finally, reconstruction from peak lists was discussed (ab initio reconstruction). This approach is based on mass shifts between masses in a peak list. If the mass shift corresponds to a biochemical transformation mass difference (e.g. methylation) then an edge is added. This definition implies the presence of a lot of false positive edges. Discussion then focused on ways to automatically filter out these edges.

4.12 Version control and CI for MS/MS libraries

Steffen Neumann (IPB – Halle, DE)

The Spectral Libraries session dealt with spectral libraries, such as GNPS, MassBank or HMDB. We discussed several routes to get spectra into libraries, ranging from conversion of existing (in-house) libraries to automated workflows extracting clean spectra from raw data. The MassBank team presented an approach to use established processes from software engineering, in particular version control and continuous testing. MassBank is moving towards a github-supported workflow in the near future, and first prototypes were shown. Such setups also simplify large-scale automatic curation, for structures, additional metadata and links. Several participants have done curation and processing of libraries, and these could then be fed back to the repositories.

4.13 False discovery rates

Marcus Ludwig (Universität Jena, DE)

It is understood that false discovery rate estimation is a necessity to allow comprehensive statistical analysis of structural identifications from mass spectrometry data. Currently no computational method provides a score to sufficiently separate true and bogus hits. Some strategies already applied in proteomics might help, such as fitting score distributions and estimating p-values. We discussed whether improvements on the measurement side might provide better data to discriminate between different candidates. However, it is unclear if mass spectrometry data provides enough information to distinguish highly similar metabolites. Due to metabolites large structural diversity and insufficient information we might need to reformulate what we can in fact deduce from the data and what we accept as a "correct" hit.

4.14 Feature prioritization

Sarah Scharfenberg (IPB – Halle, DE)

Feature prioritization is one of the main steps in a biological study although it is sparsely addressed in general MS1 workflows. Reducing the number of correlated variables in advance also positively affects the outcome of the multivariate analysis. Strongly dependent on the study goal there is a set of commonly used methods, such as PCA, PLS, variable clustering, fold changes, hypothesis testing, random forests, or a combination of these. Based on MSMS networks or visual models it is possible to connect feature intensities to bioactivities or regions. A proper experiment design enables strict criteria on how to select the relevant features, e.g. the most differentially expressed features.

Participants

- Hayley Abbiss
Murdoch University, AU
- Theodore Alexandrov
EMBL Heidelberg, DE
- Manor Askenazi
Biomedical Hosting – Arlington, US
- Eric Bach
Aalto University, FI
- Ruth Birner-Grünberger
Universität Graz, AT
- Sebastian Böcker
Universität Jena, DE
- Corey Broeckling
Colorado State University – Fort Collins, US
- Christoph Büschl
Universität für Bodenkultur – Wien, AT
- Ricardo Da Silva
University of California – San Diego, US
- Pieter Dorrestein
University of California – San Diego, US
- Edward M. Driggers
General Metabolics – Wilchester, US
- Kai Dührkop
Universität Jena, DE
- Madeleine Ernst
University of California – San Diego, US
- P. Lee Ferguson
Duke University – Durham, US
- Nils Hoffmann
ISAS – Dortmund, DE
- Julijana Ivanisevic
University of Lausanne, CH
- Fabien Jourdan
INRA-ENVT – Toulouse, FR
- Alexander Kerner
Lablicate GmbH – Hamburg, DE
- Oliver Kohlbacher
Universität Tübingen, DE
- Martin Krauss
UFZ – Leipzig, DE
- Martin Loos
looscomputing – Dübendorf, CH
- Marcus Ludwig
Universität Jena, DE
- Marnix Medema
Wageningen University, NL
- Kris Morreel
Ghent University, BE
- Rolf Müller
Helmholtz-Institut – Saarbrücken, DE
- Steffen Neumann
IPB – Halle, DE
- Louis-Felix Nothias-Scaglia
University of California – San Diego, US
- Andrew Palmer
EMBL – Heidelberg, DE
- Alan Race
Universität Bayreuth, DE
- Stacey Reinke
Murdoch University, AU
- Simon Rogers
University of Glasgow, GB
- Juho Rousu
Aalto University, FI
- Sarah Scharfenberg
IPB – Halle, DE
- Jennifer Schollee
Eawag – Dübendorf, CH
- Emma Schymanski
University of Luxembourg, LU
- Jan Stanstrup
University of Copenhagen, DK
- Michael Andrej Stravs
Eawag – Dübendorf, CH
- Raf Van de Plas
TU Delft, NL
- Justin van der Hooft
Wageningen University, NL
- Kirill Veselkov
Imperial College London, GB
- Ron Wehrens
Wageningen University, NL
- David Wishart
University of Alberta – Edmonton, CA
- Michael Anton Witting
Helmholtz Zentrum – München, DE
- Nicola Zamboni
ETH Zürich, CH