Report of Dagstuhl Seminar 15492

Computational Metabolomics

Edited by

Sebastian Böcker[1], Juho Rousu[2], and Emma Schymanski[3]

1 **Universität Jena, DE,** sebastian.boecker@uni-jena.de
2 **Aalto University, FI,** juho.rousu@aalto.fi
3 **Eawag – Dübendorf, CH,** emma.schymanski@eawag.ch

Abstract

The Dagstuhl Seminar 15492 on Computational Metabolomics brought together leading experimental (analytical chemistry and biology) and computational (computer science and bioinformatics) experts with the aim to foster the exchange of expertise needed to advance computational metabolomics. The focus was on a dynamic schedule with overview talks followed by breakout sessions, selected by the participants, covering the whole experimental-computational continuum in mass spectrometry, as well as the use of metabolomics data in applications. A general observation was that metabolomics is in the state that genomics was 20 years ago and that while the availability of data is holding back progress, several good initiatives are present. The importance of small molecules to life should be communicated properly to assist initiating a global metabolomics initiative, such as the Human Genome project. Several follow-ups were discussed, including workshops, hackathons, joint paper(s) and a new Dagstuhl Seminar in two years to follow up on this one.

Seminar November 29 to December 4, 2015 – http://www.dagstuhl.de/15492
1998 ACM Subject Classification J.3 Life and medical science
Keywords and phrases algorithms, bioinformatics, cheminformatics, computational mass spectrometry, computational metabolomics, databases
Digital Object Identifier 10.4230/DagRep.5.11.180
Edited in cooperation with Céline Brouard and Franziska Hufsky

1 Executive Summary

Sebastian Böcker
Juho Rousu
Emma Schymanski

Metabolomics has been referred to as the apogee of the omics-sciences, as it is closest to the biological phenotype. 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 spectrometers and experimental workflows during the last decades enable novel investigations of biological systems on the metabolite level. But these advances also resulted in a tremendous increase of both amount and complexity of the experimental data, such that the data processing and identification of the detected metabolites form the largest bottlenecks in high throughput analysis. Unlike proteomics, where close co-operations between experimental and computational scientists have been established over the last decade, such cooperation is still in its infancy for metabolomics.

The Dagstuhl Seminar on Computational Metabolomics brought together leading experimental and computational side experts in a dynamically-organized seminar designed to foster the exchange of expertise. Overview talks were followed by breakout sessions on topics covering the whole experimental-computational continuum in mass spectrometry.

2 Table of Contents

Executive Summary

Major topics

Generating Spectra in silico

Breakout Groups

3 Major topics

3.1 Data exchange

Pieter Dorrestein (University of California – San Diego, US)

Much discussion over the past decade in metabolomics has been around data sharing. Several metabolomics repositories exist. I asked how many people here have gone to those databases and used a dataset. Only three people raised their hands, yet this it the community that is developing tools for analysis of datasets. There are several purposes for databases:

- to capture and share metabolomics knowledge,
- to share data,
- to make chemical knowledge accessible,
- to associate metadata with the chemical knowledge.

Then one can build the computational infrastructure to retrieve metabolomics knowledge. An argument was made that we should build an analysis infrastructure that organizes and visualizes data while capturing the data metadata and computing in a distributive fashion. Future opportunities are:

- creation of living data, where data is transferred to users,
- connection to genomic information,
- assessing in silica approaches for new spectral matching functions/algorithms with a common set of LC-MS data sets (e.g. 100,000 data sets),
- relaying new information obtained with new tools to users, rather than each user doing their own search.

3.2 Searching in Structure Databases

David Wishart (University of Alberta, Edmonton, CA)

The presentation described the current state of searching for compounds in metabolic databases. There are three kinds of databases: general compound databases, public repositories and spectral databases. A major problem with the general compound databases is that they do not provide species or functional information regarding the compounds. As a result, there are now a growing number of species-specific compound databases.

This presentation also reviewed some of the key challenges facing metabolomics with regard to molecular structures searching. In particular:

- While the size of the spectral databases is growing, the actual number of compounds is not. How to increase these numbers?
- Only a small fraction of currently known metabolites have (or will have) reference LC-MS spectra. This is a real knowledge dichotomy!
- Even if we would product MS spectra for all know compounds, we would likely only identify 30% of the compounds in untargeted LC-MS. What are we missing?

I discussed some possible solutions to these, including:

- the development of compound libraries and compound exchanges,
- the development of MS/MS production tools like CFM-ID or CSI:FingerID,
- the development of structure/metabolite prediction tools.

3.3 Incorporating Experimental Knowledge

P. Lee Ferguson (Duke University, Durham, US)

Experimental knowledge can be used primarily in two ways to identify compounds in non-targeted high resolution mass spectrometry workflows. First, data such as chromatography retention time, ionization performance, and metadata such as reference count and chemical production volume can be used to refine compound identification, after data acquisition. Second, experimental data such as fates or effects of compounds can be used to prioritize data features for subsequent identification. Frontiers such as LCxLC and X-ray crystallography were introduced as future directions.

3.4 Using retention index information of an orthogonal filter for compound identification in GC/MS analysis

Tom Wenseleers (KU Leuven, BE)

In this talk I gave an overview of the potential of using retention index information for compound identification in GC/MS analysis, especially when combined with other pieces of orthogonal information, including electron impact and chemical ionization spectra, *in silico* predicted EI spectra and mass and isotope abundance information.

I provided several examples of compounds where retention index was really critical for correct identification, even if EI mass spectral fragments and mass could be measured with perfect accuracy. I then pointed out the potential of building combinatorial libraries with compounds that are biologically plausible and adding *in silico* predicted EI spectra and retention indices. A proof-of-concept was provided where this method was able to correctly identify ca. 10000 methylalkanes. I finished by discussing database requirements and the need for standardized data formats to include and more retention index information.

3.5 Utilization of retention time in LC-MS

Michael A. Witting (Helmholtz Zentrum, München, DE)

A lot of effort is made to analyze MS, MS2, MS3 … spectra, but orthogonal information like separation dimension or ion mobility are often neglected. However to improve identification of unknown or verification of known molecules they have to be incorporated. To facilitate data sharing a novel retention time indexing for RP-LC-MS was presented. This indexing system will potentially allows integrated analysis of RTI data from different sources compiled on similar systems. Additionally, *de novo* prediction of retention times using different published methods was discussed. Several limitations have been identified, which have to be tackled by the community. Lastly, ion mobility as orthogonal method was presented.

4 Generating Spectra in silico

4.1 In Silico Mass Spectral Identification

Tobias Kind (University of California – Davis, US)

In silico methods for mass spectrometry can be used to calculate spectra directly from chemical structures. Traditionally spectra had to be acquired by experimental measurements only, now purely computational methods can be used. This includes *ab initio* methods, machine learning methods, reaction based tools and heuristic methods. Their outputs have to be validated and prediction accuracy has to be tuned for better performance. In the future it will be possible to generate millions of mass spectra (hopefully highly accurate), which then will lead to the following problem: the curse of similarity and potential database poisoning with millions of similar spectra.

4.2 Competitive Fragmentation Modeling

Felicity Allen (University of Alberta, Edmonton, CA)

Existing methods for spectrum prediction generally produce far more peaks than actually occur in a measured spectrum. Competitive Fragmentation Modeling (CFM) is a method that we propose to predict fewer peaks that are more likely to occur. It uses a probabilistic, generative model of the fragmentation process. Parameters of the model are learned from data using expectation maximization. The method has recently been extended for use with EI-MS. Empirical results show that the method outperforms existing computational tools, but is still inferior to actually measuring the spectrum. Despite this short-coming, actual measurements are often costly or infeasible, and so this methods offers an important alternative.

5 Breakout Groups

5.1 Spectral Simulation

The discussions on spectral simulation started with a survey of who uses what: CFM-ID, QC (quantum chemical)-EI-MS, CSI:FingerID, Mass Frontier, ACD MS Fragmenter, HAMMER, manual interpretation, or a combination of all were mentioned. It was established that mass spectral simulation software needs to accurately predict fragment ions and their peak abundances. Most software produce different fragments and although better ranking results are achieved with e.g. CFM–ID, the fragments are not always “chemically sensible” and in this sense Mass Frontier is often more accurate because it makes use of reaction chemistry from the reference literature. The quantum chemical simulation of Grimme (QC-EI-MS) is promising and theoretically extendable to ESI but because of the complexity of the computational tasks, the quantum chemical community needs to be engaged to solve this. It was discussed whether CFM–ID could "learn" rearrangements, but it needs the knowledge in advance to do this; these cannot be exported from Mass Frontier. Toolkits used included RDkit (C++/python) ChemAxon (free academic), CDK (limited reaction capabilities) – having an active development community behind is essential. The need for more experimental data was discussed, because more data could be used to improve modelling accuracy, once large enough validation sets are available. It was debated whether the Markov approach behind CFM–ID could be used to train intensities for some of the other *in silico* fragmenters. Last ideas included treating the mass spectrum as a picture (picture recognition algorithm) and whether mass spectral data should be uploaded to http://www.kaggle.com (a platform for data prediction competitions) to get very good machine learners working on mass spectra.

5.2 Next Generation Computational Methods

The breakout group on next generation computational methods covered several topics. A debate about identification measures covered whether the current scores for *in silico* fragmenters are sufficient in separating the true from false matches and whether the score should aim to pick the best candidate or rather show how good the prediction is, also considering top K instead of top 1 (see also "Statistics", below). The "Percolator approach" was also discussed.

The next topic covered joint identification, using the presence of other substances to elevate the ranks of "unknowns" with prior evidence, using mass differences and also clustering by using multiple measurements as training sets to perform machine learning. Estimates included requiring half the number of samples for the number of metabolites under investigation (i.e. under 1000 samples for typical cases).

Finally, discussions ended with substances that are not in the databases and using predicted transformations to help find potential candidates via biotic and abiotic reactions. The presence of peptides, oligonucleotides, sugars and homologue series were also discussed, including the potential to run all small poly-peptides, potentially up to 8, and add them to the Global Natural Product Social Networking (GNPS) library. Discussions ended on a summary figure from GNPS that showed that there is a lot of "dark matter" remaining and very few known annotations, many of the unknowns are singletons.

5.3 Metadata and common input/output formats

The breakout group on metadata focused on what types of metadata would need to be reported for a given study for it to be useful and discussed resurrecting an old SepML standard using controlled vocabulary from existing ontologies. A large number of action points were made, especially involving vendors and Proteowizard, to enable export of given parameters into the open format. Points to discuss in the future remained most recent separation advances: 2D LC and GC (liquid and gas chromatography) as well as ion mobility.

The group on common input/output formats discussed the need to explore common parameters and formats between most software for small molecule identification. Two different use cases evolved: development (simple text-based format, e.g. MGF, Mascot Generic Format) versus pipeline integration once developed (fancy mzML-type format for machine-readable properties). Software-specific parameters can remain flexible. The ability of mzML to support structures may be a limitation with this format. Outputs in CSV files with common column headers or SDFs with common tags were discussed; developers should not rely on a certain order in the CSV for maximum flexibility. Some discussions on potential test data were made. These discussions will continue beyond Dagstuhl.

5.4 Integrative Omics

The breakout group on "integrative omics" discussed that the correlations between the different omics levels are complex and the integration of metabolomics is poor, with no computationally-feasible way to connect the layers. Several issues were discussed to address the lack of interaction information between metabolites and genes/proteins, such as enzyme reaction models, systematic studies of metabolite-protein binding (technically difficult to find), collation of existing knowledge in a protein–metabolite–interaction database in a machine-readable way, as well as computational methods needed to find novel pathways and interactions between different levels (text-mining?).

The combination of transcriptomics with metabolomics was discussed, rather than pure mapping, as this is more orthogonal that proteomics/metabolomics. This could be used to find the most interesting sites in the networks and possibly even help build the network if one could differentiate the data sufficiently. However, this may be hindered by different time-scales as the metabolome changes extremely fast. Finally, correlation feature-based instead of identification-based approaches were mentioned.

5.5 The Dark Matter of Metabolomics

The breakout group on the dark matter of metabolomics and in-source fragmentation phenomena had a pretty wide ranging discussion focusing on the relatively low rates of annotation of compounds/features from LC-MS studies using either MS level data, MS/MS data, or infusion data. The consensus was that 30% seems to be an approximate maximum success rate across labs. The need for a gold-standard ground truth dataset was stressed, to evaluate the various steps in the data processing and annotation processes, from peak picking/feature grouping through the final annotation and evaluation. The need for the full utilization of all existing MS data, and supplementing with non-MS data (biology, computation, NMR, etc) was reiterated to try and address the identification of real and reproducible signals.

5.6 Statistics

The statistics breakout group discussed issues that arise when searching in larger (spectral or molecular structure) databases. Currently, only relatively few compounds are identified in an LC-MS run; when more compounds are putatively identified, this will come at the price of more bogus identifications. This is independent of the fact whether we are searching in a large spectral library, or a large molecular structure database. To this end, scores have to be introduced that express a methods "confidence" that a certain identification is correct. Beyond that, False Discovery Rates (q-values, p-values) would be very helpful to navigate the putative identifications and to find reasonable thresholds of what to accept and what to reject, similar to Shotgun Proteomics. We also discussed the problem of p-value corrections for repeated testing. Finally, we discussed how to combine orthogonal information for compound identification into a single, statistically meaningful measure.

5.7 Metabolite Prediction

The metabolite prediction breakout group discussed two approaches to metabolite prediction:

1. iteratively: start from a set of known compounds, predict, confirm the existence and use this information to refine predictions
2. databases: generate predicted metabolites from large sets of known compounds and filter these "on the fly" – with the risk of combinatorial explosion

The consensus was that a combination of both approaches would be the most practical. As only a fraction of the metabolites in a metabolic network are observed, multiple prediction steps are need to be applied before a path can be confirmed, adding to the combinatorial explosion issue. On the other hand instruments are becoming more sensitive and larger fractions of (predicted) metabolites can be expected to be seen.

Big differences exist in the amount of data available in different "domains of metabolism". In some domains there is enough data to train probabilities (drugs), while in other domains data is scarce and rules are more literature based. In the case of gut transformations rules may represent what goes into a microbe and what comes out, rather than substrates and products of an enzyme. The same may be true for environmental applications.

In addition to empirical or trained likelihoods of biotransformation, kinetic parameters (from simulations) and thermodynamic parameters (which can be calculated) are useful additional parameters to evaluate and prune predicted networks.

5.8 Data visualization

The data visualization group discussed the visualization of complex data in a biological context. Interactive visualization allowing the navigation and exploration of data, going back and forth between the data and the outcomes, was a main topic. The output devices were to be "papers"/software/web apps. Another visualization challenge is looking at the large "lists" of metabolite structures, for instance the hierarchical clustering of metabolite structures in MetFragBeta, also shown in Figure 2 of Schymanski *et al.* 2014. Molecules in chemical space can also be plotted in a PCA format using chemical descriptors, as done in Figure 4 from Kuhn *et al.* 2009.

5.9 The CASMI contest

Steffen Neumann (Leibniz Institute of Plant Biochemistry – Halle, DE)

This breakout session discussed the Critical Assessment of Small Molecule Identification (CASMI) contest, founded in 2012 (http://www.casmi-contest.org). The protein equivalent, CASP, has many more participants but took several years to establish and receives considerable funding each year to run the contest. Several suggestions for future CASMIs were discussed. Participants requested raw data in addition to peak lists, with future peak lists to be provided as MGF as a new standard format for identification tools, with challenges submitted to MassBank. A "spectrum-only" category was discussed, where common candidate lists could be provided and no additional scoring criteria would be allowed, to focus on only *in silico* fragmentation techniques. A detailed description of the analytical conditions (chromatography, mass spectrometry) should be provided. The participants also indicated that they would like a CASMI workshop to discuss the results after closure of the contest; the current "outlet" is in the form of publications, with mixed success. A workshop is under consideration for the 2016 contest. Ideas for future CASMIs included a staged contest (automatic approaches first, results are then published on the website and then manual users have a few more weeks), assigning manual users a sub-category of automatic categories, to enable bigger automatic datasets for statistical robustness, and a "whole box" category where all information sources are allowed. Nuclear Magnetic Resonance spectroscopy was discussed as a new category, as there have been interesting developments recently. The idea of a GNPS/CASMI continuous evaluation dataset was also received positively and there are several challenges (unsolved) available on GNPS already.

5.10 Workflows

The workflow breakout group discussed standardized formats (see also Section 5.3) and that mzTab and mzML would be the potential file types to incorporate all information needed. Participants were strongly encouraged to pass on their ideas for standardization to the Proteomics Standards Initiative (PSI) and ask them to integrate them (and also participate in the initiative). The Spring PSI meeting (April 2016, Ghent) would be an opportunity for this. There were some additional discussions on the contents of the standards as well. Finally, although many pipelines try to get an "all in one" workflow, it was discussed about whether to split workflows into parts, with the large divide (everything before you start to work with statistics) and (after).

5.11 Feature Finding, Quantification, Labelling

Several topics merged into one breakout session. The computational challenges of quantification were discussed, including

- finding all features is challenging (needs to be more flexible/robust, e.g. slow-release substances, presence of m/z and intensity shifts, physical interferences).
- summing the signal to quantify.
- feature alignment across samples is considered essentially solved.
- still no clear idea what is the best normalization method, as this is dependent on experimental design.
- that experimental data contains no real ground truth, but while synthetic data is not appreciated by experimentalists, this is essential for computational people.
- reference datasets are available on the CompMS website.

From the experimentalists point of view, concentrations/quantification is needed to translate detected metabolites to the biology; quantification can be used to model metabolic networks and see fluxes. Instrument ionization is complex and formation of ions varies greatly with structure. Internal standards (preferably isotopically-labelled) are needed; at least one per compound class. Standard additions also possible. The solvent composition can have a huge influence on signal intensities, while the influence of acidity and polarity was also discussed. Questions included whether to sum intensities from all adducts, or remove/ignore smaller signals, how to extract response factors from runs and using ion current measurements to correct for ESI spray fluctuation. Can adduct species be predicted? Labelling experiments can yield even more information, including qualitative and quantitative flux measurements and thus tracking origin and fate of metabolites, yet over 65 % of signals remain unidentified despite labelling proving they have biological origin – see Section 5.5.

6 Hands-on Sessions

A number of small hands-on sessions were run during the meeting. The environmental and xenobiotic session on Tuesday discussed data from different sources in detail and the surprising complementarity observed in the production volume and patent data. At the same time, a breakout on the SPectraL hASH (SPLASH) introduced this concept and determined that these are now google-searchable. One participant now has a roadmap to contribute his substances to MassBank, using MetShot and RMassBank. On the last day, a software demonstration and feedback session was run across the whole morning and was enjoyed by all participants with very honest and constructive feedback and discussions about different approaches.

7 Wrap-ups

The seminar wrap-up started with expressions of interest for a commentary/perspectives paper as a partial summary of discussions – over half of the participants were interested and Pieter Dorrestein will take the lead. Focus on metabolomics and the extension to the exposome and small molecule characterization (chemical genomics? chenomics?). Michael Witting advertised a special issue about unknown identification coming up in *J. Chrom. B* (deadline mid 2016). Lee Ferguson announced the Nontarget 2016 conference in Switzerland, May 29 to June 3. A couple of new ideas such as a society for small molecule characterization or a new open source journal were considered unlikely to get off the ground, but alternative meetings such as in conjunction with the Metabolomics Society conference were considered positively. All participants indicated that they had enjoyed the meeting and would come again; none raised their hand for the opposite. The seminar wrap-up concluded with two main questions:

1. Where do we want to be in a year?
   Establishment of benchmark datasets and standard in/out data structure, improved data and spectral sharing as well as using bioboxes for modular workflows.
2. How to we encourage more people?
   Offer machine learning challenges, expose students to metabolomics, increase the data availability, improve the community building efforts (with workshops such as this Dagstuhl Seminar) and initiatives such as Computational Mass Spectrometry (CompMS), which has coursework on computational metabolomics and proteomics.

Excursion

The excursion on Wednesday afternoon was to Trier, including a city tour and the Christmas market, before dinner near the cathedral. A good time was had by all.

8 Conclusion

The first Dagstuhl Seminar on Computational Metabolomics was a huge success with positive feedback from all participants. A general observation was that metabolomics is in the state that genomics was 20 years ago and that while the availability of data is holding back progress, several good initiatives are present. The importance of small molecules to life should be communicated properly to assist initiating a global metabolomics initiative, such as the Human Genome project. Several follow-ups were discussed, including workshops, hackathons, joint paper(s) and a new Dagstuhl seminar in two years similar to this one.

The organizers wish to acknowledge the contributions of Tobias Kind, who attended on behalf of Oliver Fiehn, Franziska Hufsky and Céline Brouard who collected and typed the hand-written abstracts as well as all participants for their contributions.

Participants

- Felicity Allen
University of Alberta – Edmonton, CA
- Nuno Bandeira
University of California – San Diego, US
- Sebastian Böcker
Universität Jena, DE
- Corey Broeckling
Colorado State University – Fort Collins, US
- Céline Brouard
Aalto University – Espoo, FI
- Jacques Corbeil
University Laval – Québec, CA
- Pieter Dorrestein
University of California – San Diego, US
- Kai Dührkop
Universität Jena, DE
- P. Lee Ferguson
Duke University – Durham, US
- Franziska Hufsky
Universität Jena, DE
- Gabi Kastenmüller
Helmholtz Zentrum – München, DE
- Tobias Kind
Univ. of California – Davis, US
- Oliver Kohlbacher
Universität Tübingen, DE
- Daniel Krug
Helmholtz-Institut, DE
- Kris Morreel
Ghent University, BE
- Steffen Neumann
IPB – Halle, DE
- Tomas Pluskal
Whitehead Institute – Cambridge, US
- Lars Ridder
Netherlands eScience Center – Amsterdam, NL
- Simon Rogers
University of Glasgow, GB
- Juho Rousu
Aalto University – Espoo, FI
- Emma Schymanski
Eawag – Dübendorf, CH
- Huibin Shen
Aalto University – Espoo, FI
- Christoph Steinbeck
European Bioinformatics Institute – Cambridge, GB
- Michael Stravs
Eawag – Dübendorf, CH
- Ales Svatos
MPI für chemische Ökologie – Jena, DE
- Tom Wenseleers
KU Leuven, BE
- Rohan Williams
National Univ. of Singapore, SG
- David Wishart
University of Alberta – Edmonton, CA
- Michael Anton Witting
Helmholtz Zentrum – München, DE
- Gert Wohlgemuth
University of California – Davis, US
- Nicola Zamboni
ETH Zürich, CH