Report from Dagstuhl Seminar 22181

Computational Metabolomics: From Spectra to Knowledge

Corey Broeckling[*1], Timothy Ebbels[*2], Ewy Mathé[*3], Nicola Zamboni[*4], and Cecilia Wieder[†5]

1 **Colorado State University – Fort Collins, US.** corey.broeckling@colostate.edu
2 **Imperial College London, GB.** t.ebbels@imperial.ac.uk
3 **National Institutes of Health – Bethesda, US.** ewy.mathe@nih.gov
4 **ETH Zürich, CH.** zamboni@imsb.biol.ethz.ch
5 **Imperial College London, UK.** cecilia.wieder19@imperial.ac.uk

Abstract

The fourth edition of the Computational Metabolomics seminars, Dagstuhl Seminar 22181, brought together a wide range of computational and experimental experts to share state-of-the art methodologies and push our collective understanding of how to interpret and maximise insight of metabolomic data. With increasing amounts of metabolomic data being generated, including large-scale epidemiological studies, and increasing sensitivity of instrumentation, development of sophisticated and robust computational solutions is required. Further, community agreement on which data standards should be used and which data sets are most apt for benchmarking computational tools is needed in the field. Building upon the previous successful formats of previous seminars (17491, 15492, and 20051) on this topic, attendees gathered each morning to collectively agree on the number of sessions and topics to discuss. A summary of the daily sessions were shared amongst all participants after dinner during each day's final formal session. Further, informal evening sessions were spontaneously created to further dive into specific topics. As with past seminars, this format was very well received and enabled all participants to weigh in. Of particular note, this seminar was delayed and travel was complicated due to the pandemic. Despite these setbacks, this seminar brought together a balanced number of previous and new, seasoned and early career participants. All participants were active in these discussions, and a true sense of renewed energy ensued from the seminar. This report provides highlights of formal and informal evening sessions, including future anticipated research directions rooted from this seminar. Possible future workshops, such as a next phase of this Computational Metabolomics Dagstuhl seminar in late 2023 or 2024 were also discussed and will be applied for.

Seminar May 1–6, 2022 – http://www.dagstuhl.de/22181
2012 ACM Subject Classification Applied computing → Life and medical sciences
Keywords and phrases bioinformatics, cheminformatics, computational mass spectrometry, metabolite identification, computational metabolomics, machine learning, data integration, pathway analysis
Digital Object Identifier 10.4230/DagRep.12.5.1

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Computational Metabolomics: From Spectra to Knowledge, *Dagstuhl Reports*, Vol. 12, Issue 5, pp. 1–16
Editors: Corey Broeckling, Timothy Ebbels, Ewy Mathé, and Nicola Zamboni

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

1 Executive Summary

Corey Broeckling (Colorado State University – Fort Collins, US
Timothy Ebbels (Imperial College London, GB)
Ewy Mathé (National Institutes of Health – Bethesda, US)
Nicola Zamboni (ETH Zürich, CH)

Metabolomics is the study of small molecules in living systems, including those which generate the energy to sustain life, those that form the building blocks of macromolecules such as DNA, as well as some originating outside the living system such as pollutants. Biologically, this field is of increasing importance due to its strong connection to organism function. Metabolomics is rapidly expanding with significant advances in both measurement technology (e.g. mass spectrometry, chromatography, NMR spectroscopy) and informatics approaches. The amount and complexity of data routinely exceeds the capacity of typical software and other computational systems used in bioanalytical labs and there is an ongoing and increasingly acute need for improvements in computational, informatics and statistical/machine learning approaches to make sense of it all.

This seminar, the fourth in the series on computational metabolomics, continued some themes previously well developed, and explored many new ones. A good example of the former is the problem of how to use mass spectral data to annotate (putatively identify) the 1000s of unknown metabolites typically observed in routine assays. Another example would be the discussion of new developments in dealing with Data Independent Acquisition which has diversified considerably in the last 5 years. Many new directions were also discussed. For instance, the question of pathway analysis – how to generate semi-automated interpretation of metabolomics data on the level of groups of molecules working together in biological processes – is becoming more prominent as larger annotated datasets become available. Another new direction was "metaboproteomics", looking at the diverse array of interactions between metabolites and proteins, in particular in how metabolite derived post-translational modifications of proteins can be picked up in annotation pipelines. Other discussions focused on software aspects such as visualization of chemical space (a key problem in designing effective software tools) and the generation/curation of high quality data for benchmarking new informatics algorithms. A session on extended metabolic models looked at ways to link data and prediction tools from protein function studies to metabolites in order to gain new knowledge of unknown metabolic pathways. From a data generation technology perspective, while mass spectrometry (MS) dominated as expected (e.g. sessions on MS spectral quality requirements, fragmentation trees etc), the seminar extended beyond previous ones into a discussion of NMR data processing and modeling. Open databases, repositories and knowledge representation also featured their own discussions including Wikidata, CxSMILEs and Wikipathways/RaMP-DB. Finally the important issue of integrating metabolomic data with other relevant data types (e.g. genomics, proteomics etc) was discussed.

The seminar organization followed a similar flexible format to the previous one, where topics were both suggested in advance and brainstormed on the Monday. The whole group participated in brainstorming and prioritization and this was further refined each morning of the meeting. Parallel discussions were organized with the aim to minimize clashes in individual interests and at the end of each morning/afternoon session a plenary feedback session was held to disseminate the main discussion points to the whole group. Evening sessions were very popular and covered a wide range of topics. Overall the seminar was felt to be one of the most successful yet, highlighting the growing importance of computational metabolomics as a field in its own right and emphasizing the need for further meetings to address the important problems in this exciting area of research.

2 Table of Contents

3 Overview of Talks

3.1 How to use unlabeled mass spectra for machine learning

Sebastian Böcker (Universität Jena, DE)

The chemical space of molecules of biological interest is very large, and we have the problem that labelled datasets only span a fraction of this space; but more importantly regions of this space are not covered at all. The current trend is to use transformers for this purpose, and there are already several tools available. We discussed approaches that were brought up in other areas (natural language processing) but we also discussed the importance of not repeating what other fields have done. We discussed using fragmentation trees to a) annotate peaks, allowing us to work with molecular formulas instead of masses, b) data augmentation, removing subtrees, and c) for generating decoys, we also discussed different representations of molecular structures, and the merits and dangers of providing information on collision energies.

3.2 Pathway analysis in metabolomics

Timothy Ebbels (Imperial College London, UK)

The session started by collecting the different definitions of network and pathway analysis. One of the approaches is statistical analysis of sets of metabolites against a (difficult-to-come-by) background distribution, noting that a reference metabolome does not exist. Generally, a pathway is defined as a set/collection of molecules, and statistical analyses include over-representation analysis, set enrichment analysis, and topology-based approaches. The input to these approaches is typically two-fold: 1) molecules of interest, and optionally a metric that defines their relevance (e.g. p-value, effect size); 2) prior annotations (e.g. biological, chemical) from knowledge sources associated with these molecules. The session focused on over-representation and set enrichment-based approaches.

Several broad categories of challenges/questions were consistently brought up: 1) ability to apply broad methods, including those developed in genomics, to metabolomics without tailoring to specific peculiarities of metabolomics data; 2) metabolite identification and annotation (structural resolution); 3) ability to map metabolites to pathways and coverage of metabolite annotations; 4) ability to develop tools that are not strongly affected by missing annotations and different levels of uncertainty; 5) difficulty in defining a common language across communities and how this affects the utility of knowledge sources across different fields (e.g. human diseases, natural products, etc.). Further, the notion of single-sample pathway analysis was only mentioned briefly, but could hold great potential to move from univariate biomarker-type analysis towards more biochemically informed metabolic state analysis.

Despite the title “Pathway analysis in metabolomics/lipidomics and multi-omics” the session focused on the pathway and network analysis, and the multi-omics aspect was left for a later occasion.

3.3 Benchmarking data for metabolomics

Ewy Mathé (National Institutes of Health – Bethesda, US)

Reference samples are critical to evaluate reproducibility of metabolomic measurements across different platforms or laboratories. Publicly available reference datasets are useful to develop, test, and benchmark new methods and tools. The goal of this informal evening session was to survey the various reference samples and datasets that are used by the metabolomics community and to identify gaps in availability of these samples/datasets.

Several key reference samples include NIST 1950, COMETS reference samples, and MetQual, that are used to evaluate comparability of metabolite measurements across different groups/labs. Other efforts, e.g. Metabolomics Workbench and CASMI provide available key datasets that can be used to develop algorithms, including those that identify metabolites. Synthetic and simulated datasets are also being used.

There was general consensus that it is difficult to find useful benchmarking datasets, and that benchmarking data used to develop algorithms often do not publish the benchmarking data or the process used to generate the data. Possible solutions here include appropriate tags to datasets so that they could be searchable, or to create a repo of benchmarking datasets. Also the consensus is that more benchmarking datasets should be made available to cover various use cases (e.g. time-series, categorical/continuous outcomes, identification, normalization methodologies), and these should include commonly used platforms. Other gaps identified in this area include the scarcity of resources that bridge multi-omic data and the creation of a DREAM-like challenge.

3.4 Extended metabolic models

Juho Rousu (Aalto University, FI)

The aim of the session was to discuss the ways to link data and prediction tools from protein (enzyme) function studies and metabolomics in order to gain new knowledge of unknown metabolic pathways. Integration of enzyme function data with MS/MS and NMR measurements from untargeted metabolomics experiments is recognized as a field of great potential, for example, for enhancing structural annotation accuracy of unknown metabolites. Given the complexity of the overall task, it is important to identify subtasks that are solvable given the current state-of-the-art.

The discussions covered several aspects that make the prediction of enzymatic reactions in a fine-grained resolution challenging and approaches to overcome these aspects. In particular, the topology of enzyme function spaces is challenging due to the activity cliffs that make interpolation in the function space hard. Other key challenges include the scarcity and poor findability of data on the substrate specificity, promiscuity, and mode of action of enzymes as well as the lack of suitable representations of promiscuous enzyme function. In addition, the current deep learning approaches are likely not sufficient to cover the needs of enzyme function prediction at the fine-grained level. It was concluded that work is required on several fronts. First, writing a review article on the state of the art computational approaches

(especially deep learning), on enzyme function prediction in metabolomics was seen as important. The review should aim to provide a common vocabulary and collect information on data and software availability. Secondly, development of graph neural networks tailored to the enzyme-function prediction task, was identified as an important task. Thirdly, efforts will be made to collect useful software packages for tackling the problem.

3.5 Quality control in untargeted metabolomics

María Eugenia Monge (CIBION – Buenos Aires, AR)

The session covered various aspects and levels of quality control (QC) along the metabolomics workflow. A number of QC metrics can be calculated on the raw spectral data files. Several implementations exist, and can be calculated retrospectively, but also during the measurement process. This allows an instrumental- and lab health dashboard visualization and alerting, and is a first step to having a digital twin for the physical setup. Good experimental design includes multiple runs of standard reference materials and pooled sample QC. These are commonly used in the data analysis stages at the end, to ensure and sometimes correct data in the matrix prior or as part of the statistical analysis.

3.6 NMR computational approaches in Metabolomics

Panteleimon Takis (Imperial College London, UK)

NMR spectroscopy is one of the dominant analytical approaches for the deconvolution of complex matrices such as biofluids, which is widely used in the field of metabolomics with the caveat of low sensitivity. Consequently, there are plenty of challenges to overcome with computational approaches to interpret and extract as much metabolite information as possible from the 1H-NMR profile of complex mixtures. In this session, a detailed introduction to NMR and its application to biofluids was discussed. In addition, a brief explanation of what could be observed in a typical 1H-NMR profile of serum/plasma samples was discussed along with the basics of NMR.

3.7 Wikidata: empowering metabolomics research

Egon Willighagen (Maastricht University, NL)
Adriano Rutz (University of Geneva, CH)
Ewy Mathé (National Institutes of Health – Bethesda, US)

Joint work of Egon Willighagen, Adriano Rutz, Ewy Mathé

Wikidata (`https://wikidata.org/`) was founded in 2012 as a platform for collaborative data collection [1]. Similar to other wikis like Wikipedia, it crowdsources knowledge, but unlike Wikipedia, this new project is data-driven not narrative-driven. Another difference is the liberal license/waiver, CCZero, which mimics the concept of public domain. This allows many research domains across the world to adopt Wikidata to easily collect, curate, and integrate data, including the life sciences [2]. Because Wikidata has application programming interfaces to add, update (e.g. curate), and retrieve data many tools have been developed around it which empowers even more use cases. Notable for data editing are Mix'n'match, QuickStatements, OpenRefine, and Author Disambiguator. For visualization the Wikidata Query Service (WDQS, `https://query.wikidata.org/`) took Wikidata to a new level, empowered by a continuously updated SPARQL endpoint, which now hosts over 13 billion RDF triples. Based on the ideas of the WDQS many tools have been developed, including Scholia (`https://scholia.toolforge.org/`) which visualizes data from Wikidata using SPARQL queries [3] around so-called aspects for genes, proteins, metabolites, but also researchers, institutes, and journals and publishers. Scholia encourages knowledge discovery. In this session, we discussed how Wikidata organizes the development of the data in Wikidata and we discussed its mechanisms to ensure quality including the notion of collaborating crowds around WikiProject, like the one for Chemistry (`https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Chemistry`). For metabolomics research Wikidata certainly has potential. It already includes 1.3 million common chemicals (see `https://scholia.toolforge.org/chemical/`) and natural products [4] most with external identifiers to compound databases like PubChem, EPA CompTox dashboard, KEGG, HMDB, ChEBI, Chemical Abstracts' Common Chemistry, LIPID MAPS, and SwissLipids. But it also includes links to databases with experimental data, like MassBank, NMRShiftDB, the PDB database, or to the primary literature. Databases that have been using Wikidata include WikiPathways [1], Reactome, Complex Portal [6], and LOTUS [4].

References

1 Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 1063–1064, New York, NY, USA, 2012. Association for Computing Machinery.
2 Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, Sarah M Keating, Magnus Manske, Michael Mayers, Daniel Mietchen, Elvira Mitraka, Alexander R Pico, Timothy Putman, Anders Riutta, Nuria Queralt-Rosinach, Lynn M Schriml, Thomas Shafee, Denise Slenter, Ralf Stephan, Katherine Thornton, Ginger Tsueng, Roger Tu, Sabah Ul-Hasan, Egon Willighagen, Chunlei Wu, and Andrew I Su. Science forum: Wikidata as a knowledge graph for the life sciences. *eLife*, 9:e52614, mar 2020.
3 Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. Scholia, scientometrics and wikidata. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio

Ciravegna, and Olaf Hartig, editors, *The Semantic Web: ESWC 2017 Satellite Events*, pages 237–259, Cham, 2017. Springer International Publishing.

4 Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, Christoph Steinbeck, Guido F Pauli, Jean-Luc Wolfender, Jonathan Bisson, and Pierre-Marie Allard. The lotus initiative for open knowledge management in natural products research. *eLife*, 11:e70780, may 2022.

5 Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, 11 2020.

6 Birgit H M Meldal, Livia Perfetto, Colin Combe, Tiago Lubiana, João Vitor Ferreira Cavalcante, Hema Bye-A-Jee, Andra Waagmeester, Noemi del Toro, Anjali Shrivastava, Elisabeth Barrera, Edith Wong, Bernhard Mlecnik, Gabriela Bindea, Kalpana Panneerselvam, Egon Willighagen, Juri Rappsilber, Pablo Porras, Henning Hermjakob, and Sandra Orchard. Complex Portal 2022: new curation frontiers. *Nucleic Acids Research*, 50(D1):D578–D586, 10 2021.

3.8 Visualization and graphical user interfaces

Carolin Huber (UFZ – Leipzig, DE)

Visualization for MS data analysis is necessary for quality control and evaluation steps and it is preferred outside of vendor software applications. Different approaches for visualization and their advantages and disadvantages were discussed. Especially for MS scientists and coworkers without previous programming skills, the use of these tools needs to be as easily accessible as possible. As a final step, we discussed different workflow frameworks beyond the command line, and how these can also integrate more interactive visualizations.

3.9 Metaboproteomics

Lennart Martens (Ghent University, BE)

The advent of open modification search engines in proteomics, which allow the discovery of arbitrarily modified peptides from shotgun proteomics data sets, has enabled a new view on the proteome. And while specificity of these analyses at first remained an issue, the use of powerful predictors of peptide behaviour in liquid chromatography and tandem mass spectrometry (e.g., using DeepLC [1] and MS2PIP [2]) has now led to a new type of search engine that is entirely machine learning driven (notably ionbot: `https://ionbot.cloud`, [3]) and that can utilize these predictors to increase both sensitivity and specificity. By applying ionbot to a billion spectra from human samples, and 600 million spectra from mouse samples, both obtained from the PRIDE database [4], the most comprehensive map so far of

proteome-wide protein modification is obtained for these two organisms. Upon inspection of the most commonly encountered modifications, several known artefacts of sample processing are encountered (e.g., carbamidomethylation, oxidation), alongside modifications of known biological relevance (e.g., phosphorylation), but also modifications that appear tightly linked to basic metabolism, such as certain acylations. This session discussed these findings, and looked at some of the modifications and ways to analyse these in a broader context, notably their mapping onto pathways (e.g., WikiPathways [5]), and the potential effect of the exposome on proteins as measured by the resulting induced protein modifications, coining the tentative term "metaboproteomics" to describe this potential new interface between the metabolomics and proteomics fields.

References

1 Robbin Bouwmeester, Ralf Gabriels, Niels Hulstaert, Lennart Martens, and Sven Degroeve. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, 18(11):1363–1369, 2021.

2 Ralf Gabriels, Lennart Martens, and Sven Degroeve. Updated MS²PIP web server delivers fast and accurate MS² peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research*, 47(W1):W295–W299, 04 2019.

3 Sven Degroeve, Ralf Gabriels, Kevin Velghe, Robbin Bouwmeester, Natalia Tichshenko, and Lennart Martens. ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv*, 2022.

4 Lennart Martens, Henning Hermjakob, Philip Jones, Marcin Adamski, Chris Taylor, David States, Kris Gevaert, Joël Vandekerckhove, and Rolf Apweiler. PRIDE: The proteomics identifications database. *PROTEOMICS*, 5(13):3537–3545, 2005.

5 Ryan A Miller, Martina Kutmon, Anwesha Bohler, Andra Waagmeester, Chris T Evelo, and Egon L Willighagen. Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions. *PLOS ONE*, 17(4):1–21, 2022.

3.10 Data independent analysis (DIA)

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

Data independent analysis (DIA) is an MS/MS acquisition approach by which fragmentation is performed for many precursors simultaneously; the benefit of this approach is that more (all) precursors can be sampled. Some approaches may sacrifice sensitivity, and all methods will sacrifice selectivity to enable broad sampling. Discussion revolved around these trade-offs, and how DIA is being used and perceived across the community. The data analysis challenges were discussed, noting that target applications are generally a solved problem, but untargeted analysis remains challenging. The various flavors of DIA were summarized, including ion mobility spectrometry as a component of the process. Plans to develop R-based tools for a more "universal" DIA product ion assignment were outlined.

3.11 Visualization of chemical space

Justin van der Hooft (Wageningen University, NL)
Rui Pinto (Imperial College London, UK)

Joint work of Justin van der Hooft, Rui Pinto

This session started by asking what chemical space actually means. Naturally, depending on the context, this can mean different things: all known chemicals, all chemicals identified in one experiment or in a number of samples, etc. Furthermore, chemical space can be built based on structural (fingerprints) and analytical data (mass spectral similarity).

Following this, Rui Pinto demonstrated his approach to embedding metabolite features based on MS1 feature correlations. Rui showed how metabolites could be mapped onto a UMAP embedding, and how chemical compound classes were separated. Then, Justin van der Hooft showed various ways to visualize mass spectral embeddings, using treemap, t-MAP and t-SNE. Also, a 3-dimensional molecule app was demonstrated to explore the mass spectral embedding.

The group finished the session with a brainstorm on how to connect molecules using chemical and biological information. We identified many types and realized that both mass spectral similarity scores and correlation distances could be used to construct a "base network" or embedding on which other information is then mapped. This is an interesting area for future research.

3.12 MS/MS spectral quality (part 1)

Michael Andrej Stravs (Eawag – Dübendorf, CH)

Multiple topics concerning fragmentation spectra and their computational analysis were discussed. Initial discussion reflected critically a recent publication in the field proposing entropy metrics as a measure for spectral quality and spectral similarity. On a closer inspection, the metric penalizes evenness without a stringent justification. Further discussion sought to clarify the meaning of "spectral quality", as a measure to express discriminatory power of a spectrum versus a measure of cleanness i.e. freedom from artefacts/noise.

Combination (merging) and combined acquisition (ramping/stepping) of spectra is anecdotally said to enhance separation metrics, but it remains unclear whether experimental (ramping/ stepping) combined spectra fulfill the promise expected from ideal, computationally combined (merged) spectra due to loss of signal. Evaluations are planned to shed light on the matter.

Mass range: Combinatorially, we expect low mass and high-mass ions to be uninformative, due to limited combinatorial space. However in practice low-mass ions provide an information rich structural fingerprint as evidenced by machine learning methods. Weighting curves based on mass and entropy expectations have shown little use; a curve only penalizing high mass is worth trying. Further, the "quality" of a spectrum could be approximated by how well a fragmentation tree, its model, can represent it. We note that reproducible unannotated peaks still are frequent in multiple libraries. Finally, correlation between scans across an entire dataset (as opposed to correlations in a single measurement) show promise for fragment deconvolution both for targeted and broad-range fragmentation.

3.13 MS/MS spectral quality (part 2)

Adriano Rutz (University of Geneva, CH)

A visualization tool to explore possibly unexplained fragments (i.e. no molecular formula but possibly good intensity correlation with other peaks) was showcased. This tool was presented along with a method to minimize overlap when acquiring reference spectra to build libraries. The influence of this minimized overlap could help better explain (or minimize) the previously unexplained fragments. Multiple origins of these unexplained fragments were discussed, such as complex multi-charged multimers, or fragments originating from instrumental issues such as "peak ringing" for ToF spectrometers.

A need for a new metric enabling the estimation of the percentage of fragmentation independently from the collision energy was discussed. This will be a great help for further machine learning applications. The highest obstacle to the implementation of such a metric resides in the "left-censored" spectra (e.g. acquisition starting at 50 m/z). Ways to implement such a metric and overcome this limitation were discussed and cumulative distribution of intensities from precursor down was proposed.

Finally, some new fragmentation methods (e.g. EAD, UVPD) supposed to improve spectral quality were discussed. While some of them seem promising, there is actually not enough data fully evaluated to draw any conclusions at this time.

3.14 CxSMILES: computation ready representation for compound classes

Egon Willighagen (Maastricht University, NL)
Adriano Rutz (University of Geneva, CH

Joint work of Egon Willighagen, Adriano Rutz

The 2017 Dagstuhl Seminar 17491 [1] meeting posed the problem that experimental characterization of measured metabolites often has an uncertainty in the annotation of their chemical identity. However, computational metabolomics often requires a computer representation to be able to do calculations with the chemical structures. During the 2020 Dagstuhl Seminar 20051 [2] meeting cheminformatics was explored in more detail. ChemAxon Extended SMILES (CxSMILES) and Markush Structures were identified as possible solutions to represent compound classes.

In this 1.5 hour session a group of seven Seminar participants continued to explore the power and limitations of CxSMILES and continued working on a write-up to explore the use and cheminformatics implementation in the Chemistry Development Kit [3], online available at `https://egonw.github.io/cdk-cxsmiles/`. Particularly, we added a new "Classes of compounds and where to find a CxSMILES" section with example CxSMILES for a number of compound classes. Another notable improvement during this Dagstuhl Seminar was that the CxSMILES property P10718 [4] in Wikidata was accepted, allowing us to archive and disseminate CxSMILES representations in Wikidata. Enumeration of CxSMILES, their integration into chemoinformatic pipelines for properties calculations, or conversion to InChI(-Keys) remains to be explored.

References

1 Theodore Alexandrov, Sebastian Böcker, Pieter Dorrestein, and Emma Schymanski. Computational Metabolomics: Identification, Interpretation, Imaging (Dagstuhl Seminar 17491). *Dagstuhl Reports*, 7(12):1–17, 2018.
2 Sebastian Böcker, Corey Broeckling, Emma Schymanski, and Nicola Zamboni. Computational Metabolomics: From Cheminformatics to Machine Learning (Dagstuhl Seminar 20051). *Dagstuhl Reports*, 10(1):144–159, 2020.
3 Egon L Willighagen, John W Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliazkova, Stefan Kuhn, Tomáš Pluskal, Miquel Rojas-Chertó, Ola Spjuth, Gilleain Torrance, Chris T Evelo, Rajarshi Guha, and Christoph Steinbeck. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9(1):33, 2017.
4 CXSMILES (P10718).

3.15 Estimating concentration from untargeted MS data

Anneli Kruve (Stockholm University, SE)
Steffen Neumann (IPB – Halle, DE)

Joint work of Anneli Kruve, Steffen Neumann

Currently mostly relative quantification of untargeted mass spectrometry data is used on e.g. follow up studies with identification and quantification. In relative quantification isotopically labeled C13 for cell cultures, or CO2 for plants is a good measure. Absolute quantification is used for giving results which are comparable between peaks/chemicals with different structures. For this, structurally similar standards are used but these can yield very inaccurate results due to very different behavior in MS. There are efforts to model this, but more data is needed e.g. calibration graphs. This highlights the importance of data exchange. It was also discussed that NMR is quantitative, and there have been NMR-guided MS approaches. This might open up possibilities to create quantification curves for chemicals which cannot be quantified in any other way.

3.16 WikiPathways and RaMP-DB

Egon Willighagen (Maastricht University, NL)
Ewy Mathé (National Institutes of Health – Bethesda, US)

Joint work of Egon Willighagen, Ewy Mathé

WikiPathways is an expert-curated, collaborative biological pathway database [1]. It describes biological processes for multiple species at many levels. During the meeting we discussed various aspects of the database: data format (GPML), export formats (RDF, PNG, SBML, etc), data curation, automated testing (Jenkins), and how it links to other knowledge bases (via identifier mapping databases using BridgeDb [2]).

WikiPathways is also used by other databases, e.g. for pathway enrichment and data visualization. RaMP-DB is one of those tools [3, 4]. RaMP-DB aggregates multiple types of annotations on human metabolites and proteins/genes from multiple sources to provide

a comprehensive and up-to-date source of biological and chemical annotations. RaMP-DB provides user-friendly means of interacting with the database for batch queries and enrichment analyses.

Through the session, thoughts on future maintenance and development of these resources were brought up, including the need for standards of reporting to ease harmonization across resources, assessment of the quality of annotations, and the ability to use these resources as benchmarking data for method development.

References

1 Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, 11 2020.

2 Martijn P van Iersel, Alexander R Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, Kristina Hanspers, Bruce R Conklin, and Chris T Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1):5, 2010.

3 Bofei Zhang, Senyang Hu, Elizabeth Baskin, Andrew Patt, Jalal K. Siddiqui, and Ewy A. Mathé. RaMP: A comprehensive relational database of metabolomics pathways for pathway enrichment analysis of genes and metabolites. *Metabolites*, 8(1), mar 2018.

4 John Braisted, Andrew Patt, Cole Tindall, Tara Eicher, Timothy Sheils, Jorge Neyra, Kyle Spencer, and Ewy A Mathé. RaMP-DB 2.0: a renovated knowledgebase for deriving biological and chemical insight from genes, proteins, and metabolites. *bioRxiv*, 2022.

3.17 Data collection considerations for MS^n spectral libraries

Tomáš Pluskal (The Czech Academy of Sciences – Prague, CZ)

We discussed many issues and parameters to consider when building a multi-stage MSn spectral library, such as ionization modes, chromatography conditions, scanning mass resolution, and acquiring preliminary data to inform and optimize data collection. The flow injection method was debated in more detail to extend the peak width for obtaining more fragmentation experiments per compound. Here, applying a second LC pump was mentioned as a way to add an additional makeup flow, which broadens the peak width as the first pump delivers the sample with a minimal flow rate. Furthermore, we discussed the optimization of the parameters of the MSn data acquisition method. The first parameter was the mass resolution and we discussed the community's expectations for high-quality fragmentation tree libraries. A lower mass resolution enables fast scanning but loses accuracy. Accordingly, a higher mass resolution needs more analysis time. As a compromise, the resolution of 30000 for MS1 and 15000 for MSn were suggested as a compromise, because the filling time is the more crucial step to enhance signal intensities for deeper MSn spectra. Since MS2 data are already available for many compounds in the library that we are planning to acquire, these MS2 scans should be limited, spending more time on MS3, MS4, and maybe MS5 experiments. Furthermore, the existing MS2 data can be used to plan the optimal strategy of MSn data collection for each compound. The depth and breadth of the spectral trees should be evaluated for multiple compounds as proof of principle and to avoid empty scans.

In this context, the application of multiple collision energies for one precursor needs to be examined for its usefulness. As a consequence, we agreed that the first data of hundreds of different compounds are needed to evaluate the multi-step fragmentation method.

3.18 RT, adduct formation, and calibration curve sharing

Michael Witting (Helmholtz Zentrum München, DE)

We have discussed GitHub repositories for sharing different types of data, including RT, adduct formation, and calibration curves. This data will make it possible to develop new machine learning methods for the prediction of RT, sodium and other adduct formation as well as ionization efficacy. The repository can be found at `https://github.com/michaelwitting/RtPredTrainingData`.

3.19 Metabolomics data integration

Justin van der Hooft (Wageningen University, NL)

The session started with a collection of scenarios and approaches used by the discussion group that include metabolomics-based data integration. Based on the group's input, it turned out that mostly genomics data is linked to metabolomics data. The main aims mentioned are to find biomarkers in clinical metabolomics or link genes to molecules in natural product processing. Sometimes these integrative analyses are further enriched by the addition of proteome or transcriptome data to assist in forming a systems-wide view of the data. If clinical or phenotype data is available, predictive analysis is often targeted, for example in the form of MWAS. Common challenges the group identified include the available covariates (samples), the number of false positives in correlation analysis, and the biological interpretation (of data links). The group identified that we usually link multiple omics using data tables/feature tables, which feed into correlation analysis or machine learning/multivariate statistical approaches. To make links between molecules, it is important to use consistent identifiers, and we discussed FAIR integration of data.

We concluded the session with defining the various levels of data integration: early (concatenate matrices, then model), intermediate (statistical approach with multiple input matrices to make predictions), and late (model each matrix separately and combine predictions). Finally, several dimensions of integration were discussed: across omics, within omics (e.g. different metabolomics assays), and across organisms.

Participants

- Sebastian Böcker
Universität Jena, DE
- Corey Broeckling
Colorado State University – Fort Collins, US
- Roman Bushuiev
The Czech Academy of Sciences – Prague, CZ
- Timothy Ebbels
Imperial College London, GB
- Soha Hassoun
Tufts University – Medford, US
- Carolin Huber
UFZ – Leipzig, DE
- Katerina Kechris
University of Colorado – Aurora, US
- Oliver Kohlbacher
Universität Tübingen, DE
- Anneli Kruve
Stockholm University, SE
- Tytus Mak
NIST – Gaithersburg, US
- Lennart Martens
Ghent University, BE
- Ewy Mathé
National Institutes of Health – Bethesda, US
- María Eugenia Monge
CIBION – Buenos Aires, AR
- Steffen Neumann
IPB – Halle, DE
- Louis-Felix Nothias
University of Geneva, CH
- Rui Pinto
Imperial College London and UK-Dementia Research Institute – London, GB
- Tomáš Pluskal
The Czech Academy of Sciences – Prague, CZ
- Hannes Röst
University of Toronto, CA
- Juho Rousu
Aalto University, FI
- Francesco Russo
SSI – Copenhagen, DK
- Adriano Rutz
University of Geneva, CH
- Michael Andrej Stravs
Eawag – Dübendorf, CH
- Panteleimon Takis
Imperial College London, GB
- Justin van der Hooft
Wageningen University, NL
- Cecilia Wieder
Imperial College London, GB
- Egon Willighagen
Maastricht University, NL
- Michael Anton Witting
Helmholtz Zentrum München, DE
- Mitja Zdouc
Wageningen University, NL