

Report from Dagstuhl Seminar 20051

Computational Metabolomics: From Cheminformatics to Machine Learning

Edited by

Sebastian Böcker¹, Corey Broeckling², Emma Schymanski³, and Nicola Zamboni⁴

1 Friedrich-Schiller-Universität Jena, DE, sebastian.boecker@uni-jena.de

2 Colorado State University, Fort Collins, CO, US,
corey.broeckling@colostate.edu

3 University of Luxembourg, LU, emma.schymanski@uni.lu

4 ETH Zürich, CH, zamboni@imsb.biol.ethz.ch

Abstract

Dagstuhl Seminar 20051 on Computational Metabolomics is the third edition of seminars on this topic and focused on Cheminformatics and Machine Learning. With the advent of higher precision instrumentation, application of metabolomics to a wider variety of small molecules, and ever increasing amounts of raw and processed data available, developments in cheminformatics and machine learning are sorely needed to facilitate interoperability and leverage further insights from these data. Following on from Seminars 17491 and 15492, this edition convened both experimental and computational experts, many of whom had attended the previous sessions and brought much-valued perspective to the week's proceedings and discussions. Throughout the week, participants first debated on what topics to discuss in detail, before dispersing into smaller, focused working groups for more in-depth discussions. This dynamic format was found to be most productive and ensured active engagement amongst the participants. The abstracts in this report reflect these working group discussions, in addition to summarising several informal evening sessions. Action points to follow-up on after the seminar were also discussed, including future workshops and possibly another Dagstuhl seminar in late 2021 or 2022.

Seminar January 26–31, 2020 – <http://www.dagstuhl.de/20051>

2012 ACM Subject Classification Applied computing → Life and medical sciences

Keywords and phrases bioinformatics, cheminformatics, computational mass spectrometry, computational metabolomics, machine learning

Digital Object Identifier 10.4230/DagRep.10.1.144

Edited in cooperation with Adelene Lai


1 Executive Summary

Sebastian Böcker (Friedrich-Schiller-Universität Jena, DE)

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

Emma Schymanski (University of Luxembourg, LU)

Nicola Zamboni (ETH Zürich, CH)

License  Creative Commons BY 3.0 Unported license

© Sebastian Böcker, Corey Broeckling, Emma Schymanski, and Nicola Zamboni

Mass spectrometry is the predominant analytical technique for detection, identification, and quantification in metabolomics experiments. Technological advances in mass spectrometry and experimental workflows during the last decade enabled novel investigations of biological



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Computational Metabolomics: From Cheminformatics to Machine Learning, *Dagstuhl Reports*, Vol. 10, Issue 1, pp. 144–159

Editors: Sebastian Böcker, Corey Broeckling, Emma Schymanski, and Nicola Zamboni



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

systems on the metabolite level. Metabolomics started as the study of all metabolites in a living cell or organism; in comparison to transcriptome and proteome, the metabolome is a better proxy of metabolic activity. Emerging fields including personalized medicine and exposomics have expanded the scope of metabolomics to “all” small molecules, including those of non-biological origin. Advances in instrumentation plus rapid increase in popularity, throughput and desired compound coverage has resulted in vast amounts of both raw and processed data; the field is in desperate need for further developments in computational methods. Methods established in other -omics fields are frequently not transferable to metabolomics due to the structural diversity of small molecules. This third Dagstuhl Seminar on Computational Metabolomics (following Seminars 15492 and 17491) focused on cheminformatics and machine learning. The seminar was less structured than previous seminars, forming break-out sessions already from Monday afternoon, then collecting participants back into plenary sessions at regular intervals for discussions and further topic exploration. The major topics launched on Monday included cheminformatics, genome mining and autoencoders, which were developed throughout the day. Other topics discussed throughout the week included biosynthesis and gene clusters, confidence and compound identification, spectral versus structural similarity, statistical integration, collision cross section (CCS) and ion mobility separation (IMS), benchmarking data, open feature file format, exposomics, data processing and acquisition. Several evening sessions were also held, including retention time, Bioschemas, MassBank, ethics and philosophy of software development, open biological pathways, mass spec health check, Jupyter notebooks, a mini decoy session and a session on coding tips. The excursion, breaking with previous Christmas Market traditions, was to the Völklingen steelworks. Finally, the entire seminar was wrapped up with a discussion on the future of untargeted metabolomics on Friday – time will tell what the future Computational Metabolomics Seminars will bring. A further seminar in the series may be considered for the end of 2021 or in 2022.

2 Table of Contents

Executive Summary

Sebastian Böcker, Corey Broeckling, Emma Schymanski, and Nicola Zamboni . . . 144

Break-Out Group and Plenary Discussions

Spectral vs. Structural Similarity

Oliver Alka, Adelene Lai, and Justin van der Hooft 148

Data Processing in Metabolomics

Nikiforos Alygizakis 148

MS/MS Spectrum Quality and Instrument Control

Corey Broeckling 149

Exposomics

Xiuxia Du, Kati Hanhineva, and Augustin Scalbert 149

Mass Spectrometry Coding Standards

Laurent Gatto and Ewy Mathé 150

Cheminformatics for Users

Marcus Ludwig, Steffen Neumann, and Egon Willighagen 151

The mzFeature File Format to Bridge Processing and Annotation in Untargeted Metabolomics

Tytus Mak, Oliver Alka, Sebastian Böcker, Pieter Dorrestein, Markus Fleischauer, Oliver Kohlbacher, Marcus Ludwig, Louis-Felix Nothias-Scaglia, and Tomas Pluskal 152

Benchmark Data

Ewy Mathé 152

Mining Metabolome and Genome

Ewy Mathé 153

Confidence and Compound Identification

Hunter Moseley 153

MassBank: Status Meeting

Steffen Neumann 154

Autoencoders

Jamie Nunez and Michael Andrej Stravs 154

Collision Cross Section and Ion Mobility Spectrometry

Tomas Pluskal 154

Jupyter Notebooks for #FAIR Data Science

Stacey N. Reinke 155

Statistical Integration

Stacey N. Reinke 155

Biosynthesis, Gene Clusters, and Predicting Natural Products from the Genome

Justin van der Hooft and Simon Rogers 156

Bioschemas

Egon Willighagen 156

Open Biological Pathways with WikiPathways
Egon Willighagen 157

Retention Time
Michael Anton Witting 157

Conclusion: The Future of Computational Metabolomics
Sebastian Böcker 158

Participants 159

3 Break-Out Group and Plenary Discussions

3.1 Spectral vs. Structural Similarity

Oliver Alka (Universität Tübingen, DE), Adelene Lai (University of Luxembourg, LU), and Justin van der Hooft (Wageningen University, NL)


License  Creative Commons BY 3.0 Unported license
© Oliver Alka, Adelene Lai, and Justin van der Hooft

Spectral similarity underpins many of our analyses, like the use of spectral similarity in library matching and molecular networking. This break-out group tried to reconcile spectral and structural similarity – on a fundamental level, can we equate two molecules structurally if their spectra are considered similar? Feedback collected from the group showed that cosine similarity was the most-used and perhaps well-known measure of spectral similarity because of how easy it is to calculate and wide availability in various vendor software, but that it is an imperfect measure not least because it is hard to test how it works. Further options for measuring spectral similarity discussed include Hybrid (considering fragment and losses), All Mass Differences, and performing both Forward and Reverse comparisons. The impact of different instruments and their respective vendors and options (e.g. ramped collision energy, stepped) on spectra was also discussed, with some suggestions to merge or derive an average spectrum. This could be improved using mass difference in the scoring by creating a hybrid score for example. Some concrete ideas on implementing graph-based extraction of (relevant) mass differences from spectra and using those to calculate a similarity score were also discussed.

Regarding structural similarity, Tanimoto was regarded by many as inadequate, and other methods were discussed, including fingerprint comparison, maximum common edge subgraph, and DICE. On evaluation, chemical classes predicted from spectra were proposed as alternatives to fingerprints.

3.2 Data Processing in Metabolomics

Nikiforos Alygizakis (Environmental Institute – Koš, SK)

License  Creative Commons BY 3.0 Unported license
© Nikiforos Alygizakis

Data processing pipelines consist of discrete steps (centroiding, chemical noise removal, peak picking, retention time alignment, grouping of features, componentization of isotopes, adducts and in-source fragments). Even though there is a multitude of software (both open-source and commercial) for each step of the pipeline, there is still space for improvement. Peak picking is an area with great potential for improvement and is a crucial step in metabolomics workflows. It must be highlighted that there are commercial peak pickers (e.g. Genedata Expressionists) that may also be worth implementing as open-source tools and benchmarked against established peak pickers. Little margin for improvement exists for grouping of peaks across samples and retention time alignment. Componentization and especially accurate detection of adducts in MS1 full-scan spectra is a topic that needs further investigation. Adduct formation heavily depends on the mobile phases of chromatography and physicochemical properties of the analytes. This topic has not been addressed and current mass spectral libraries rarely store MS1 spectra. Instrumental developments such as

high-resolution ($R > 500,000$) and recording of profile data motivate the need of improved componentization software that can improve annotation in metabolomics workflows. Existing software should be parallelized and new software with sophisticated computational approaches can now be applied, since computer power is readily available. Software developments should take into account the application of strong quality assurance and quality control during metabolomic experiments (e.g. QC charts, spiking of internal standards, standard operational procedures for all parts of the analysis) that needs to be implemented in all analytical laboratories. High-quality data in combination with advanced software tools can significantly improve data processing in metabolomics.

3.3 MS/MS Spectrum Quality and Instrument Control

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

License  Creative Commons BY 3.0 Unported license
© Corey Broeckling

MS/MS Spectrum quality for small molecules has historically depended on spectral similarity to library entries. Computational interpretation tools have opened the possibility to explore spectrum information content in a library independent manner. There is little rigorous description of what constitutes a high quality spectrum for small molecules, particularly in the absence of a library search. In the proteomics field, descriptors for spectrum quality have been suggested and might be adapted to metabolomics. This has yet to be experimentally and statistically determined. In general, it seems that the fragments in the middle between the minimum m/z and the precursor hold the most information, and more fragments are better than few. In addition, using different fragmentation methods, such as CID and HCD, seem not to hold additional information about quality and metabolite identification. Experimental methods offering real-time instrument control could improve the quality by using multiple collision energies or ramps, or refining collision energy on a feature-by-feature basis. The isolation window (MS1), as well as the time of sampling seem to be important. The conclusion is that spectral quality assessment needs more experimental evaluation to find valid descriptors and validate these for different experimental setups

3.4 Exposomics

Xiuxia Du (University of North Carolina – Charlotte, NC, US), Kati Hanhineva (University of Kuopio, FI), and Augustin Scalbert (IARC – Lyon, FR)

License  Creative Commons BY 3.0 Unported license
© Xiuxia Du, Kati Hanhineva, and Augustin Scalbert

The exposome encompasses all environmental exposures including chemical, physical, and biological stressors, as well as lifestyle and social environments, from conception through adulthood (<https://hhearprogram.org/>). Despite tremendous efforts that have been made by researchers in diverse areas including environmental sciences, metabolomics, nutritional sciences, etc, enormous challenges remain. One of these challenges concerns the tremendous efforts currently required to annotate exposome data. More than half of the session time was spent on discussing the causes of this challenge and potential ways to address it.

The causes include: (1) the huge chemical space that the exposome covers and further biotransformations of the compounds in this space; (2) fragmentation of available resources; (3) onerous efforts required to deposit data in repositories; (4) shortage of reference spectra for assigning spectra to compounds; (5) lack of reference exposome; and (6) shortage of training data to build automated computational tools for annotating the exposome.


This challenge can be addressed from different angles simultaneously. For example, the detected compounds can be prioritized for suspect screening based on metadata that are collected from: (1) specific experiments (e.g. curated in Metabolomics workbench, MetaboLights or GNPS), or (2) literature sources with data eventually curated in existing databases (e.g. HMDB, PubChem, FooDB, Phenol-Explorer, Exposome-Explorer).

Furthermore, additional resources and informatics capabilities would be needed to facilitate exposome annotation. These include: (1) data mining tools to collect information scattered in the literature, mainly in pdf files; (2) training data for priority scoring in annotation (e.g. CRISPR-CAS9, artificial guts, etc); (3) tools for more efficient and rapid annotation and suspect screening in metabolic profiles largely done manually so far; (4) sharing analytical/spectral data from samples and reference compounds to speed up the annotation of the exposome through a community effort; (5) better integration of different types of data from various databases (e.g. links to spectra in PubChem); and (6) resources to support deposition curation and warrant sustainability of databases.

Finally, we discussed how to further address the challenge. We asked Dr. David Wishart to lead an effort to coordinate future research and development activities by researchers. As an actionable item, Dr. Wishart will plan to host a workshop in Edmonton or the Rocky mountain parks (Canmore) in the summer of 2020.

3.5 Mass Spectrometry Coding Standards

Laurent Gatto (University of Louvain, BE) and Ewy Mathé (Ohio State University – Columbus, OH, US)

License  Creative Commons BY 3.0 Unported license
© Laurent Gatto and Ewy Mathé


During this discussion about guidelines on how to share code related to computational mass spectrometry, we decided to remain programming language agnostic, and focus on community-level goals. It was highlighted that for such contributions to be helpful, they need to contain software or code, and at least some testing data and documentation. The extent and “quality” of these elements, especially the latter, should however be regarded as flexible for two main reasons, the first one being that less seasoned contributors shouldn’t be barred from disseminating their work due to arbitrarily strict requirements. Second, there is a difference between publishing a method or a (computational) solution to a specific problem and a “finished” software product, and it is important to appreciate the value (novelty or engineering quality, for example) of both of these outputs. Hence the importance for these guidelines to emanate from the community at large to enable/facilitate important goals, and should not become rigid requirements.

We have identified three important end goals that should be highlighted when contributing and disseminating code, namely (1) reproducibility, (2) usability and (3) learnability. Each of these will require code, documentation and test data, albeit to different extents. In some cases, small test data and a README file will suffice to install and reproduce some scripts

implementing a novel method. On the other hand, finalised software products will have to provide more in-depth documentation (function-, software-level documentation, how-to's, etc.) and comply with additional (language-specific) software requirements, hence the importance for the contributors to accurately describe the type and scope of the code deliverables they share with the community.

3.6 Cheminformatics for Users

Marcus Ludwig (Friedrich-Schiller-Universität Jena, DE), Steffen Neumann (IPB – Halle, DE), and Egon Willighagen (Maastricht University, NL)

License  Creative Commons BY 3.0 Unported license
© Marcus Ludwig, Steffen Neumann, and Egon Willighagen


Cheminformatics is the use of computer and informational techniques applied to a range of problems in the field of chemistry [1]. In the context of Computational Metabolomics we represent metabolites as molecular structures, but due to the uncertainty in annotation, we need to be able to represent partially characterised structures. Representation of partial information can be distinguished into two different applications: (1) Listing the occurrence of defined substructures (fingerprint) of the measured molecule or categorizing molecules into classes, and (2) the estimation of the biggest core structure which is supported by the measured data. We concentrated on the estimation of core structures in this discussion. Since the 2017 Dagstuhl Seminar 17491 [2] methods have been developed (e.g. ChemAxon Extended SMILES (CxSMILES), Markush Structures), and examples were now created during an evening session. Discussion topics included how different layers of information provide different pieces of structural evidence, and that CxSMILES provides many solutions, but is limited. For example, for uncertainty of double bond locations in lipid tails, CxSMILES does not have a satisfactory solution. Therefore, the molecular formula and shortlists of specific compounds remain complementary. The need for open source tools to derive a common CxSMILES, depict CxSMILES, and enumerate structures starting with an CxSMILES was established. The Chemistry Development Kit is being explored for this. Another area of cheminformatics is structure generation required to identify metabolites not yet in compound databases. Existing approaches cover a continuum from unconstrained structure generation, to combinatorial decoration of frameworks or backbones and biochemical expansion of structure databases. There are cross-links to the session on autoencoders of chemical structures, which can generate structures, ideally with constraints from prior or experimental knowledge.

References

- 1 Wikipedia contributors. *Cheminformatics – Wikipedia, The Free Encyclopedia*. <https://en.wikipedia.org/w/index.php?title=Cheminformatics&oldid=909899401>, Online; accessed 3-February-2020.
- 2 Dagstuhl seminar 17491 contributors. *Computational Metabolomics: Identification, Interpretation, Imaging*. <https://www.dagstuhl.de/17491>, Online; accessed 3-February-2020.

3.7 The mzFeature File Format to Bridge Processing and Annotation in Untargeted Metabolomics


Tytus Mak (NIST – Gaithersburg, MD, US), Oliver Alka (Universität Tübingen, DE), Sebastian Böcker (Friedrich-Schiller-Universität Jena, DE), Pieter Dorrestein (University of California – San Diego, CA, US), Markus Fleischauer (Friedrich-Schiller-Universität Jena, DE), Oliver Kohlbacher (Universität Tübingen, DE), Marcus Ludwig (Friedrich-Schiller-Universität Jena, DE), Louis-Felix Nothias-Scaglia (University of California – San Diego, CA, US), and Tomas Pluskal (Whitehead Institute – Cambridge, MA, US)

License  Creative Commons BY 3.0 Unported license
© Tytus Mak, Oliver Alka, Sebastian Böcker, Pieter Dorrestein, Markus Fleischauer, Oliver Kohlbacher, Marcus Ludwig, Louis-Felix Nothias-Scaglia, and Tomas Pluskal

While there are open formats for mass spectrometry data (e.g. mzML) and downstream annotation (i.e. mzTab-M), there is currently no existing file interoperable format to bridge the gap between processing and structure annotation tools in non-targeted LC-MS/MS data processing. This proposal aims at designing an intermediate “mzFeature” open file format that would hierarchically store information on the detected spectral features that have been extracted via peak picking/feature finding algorithms (e.g. XCMS, MZmine, OpenMS). Feature objects are storing centroided spectral information (mass traces, associated MS2 spectra, MS_n etc.), along with m/z and retention time statistics (i.e. peak apex, peak start/end). These are usually extracted on a file basis and Features of the same file can be grouped as a FeatureMap. The Features can be linked into FeatureGroups across multiple mass spectrometry files, which may consist of an adduct type, isotopologues, and in-source fragments that originate from the same molecule. Ambiguities are accounted for via multiple mappings, such as an MS2 spectrum being assigned to multiple feature objects. The format should accommodate the inclusion of metadata that are specific to various instrument and processing tools.

3.8 Benchmark Data

Ewy Mathé (Ohio State University – Columbus, OH, US)

License  Creative Commons BY 3.0 Unported license
© Ewy Mathé

Benchmarking datasets are needed to test new methods. These data should be well understood and well characterised. One specific area of need is multi-omics datasets. When collecting these data, the proper meta-information (on samples and metabolites) needs to be included. There needs to be a balance between incorporating appropriate meta-information and the difficulty/time required for collecting/inputting that info.

There are multiple complementary efforts for doing this: 1) MANA SODA: community-driven input of data and software; 2) NIH Metabolomics Workbench: well curated datasets collected for benchmarking; 3) a previous Dagstuhl conference had started a similar effort for Proteomics [1]. Incentivizing data generators and software developers to submit their work (e.g. publications, advertisements, recommendations, etc.) is key to the success of these efforts. Also defining use cases, where what people want data for is defined, is important. Benchmarking data could be comparison of existing data or may require the generation of new data. The task of this session will be to define such use cases and best approaches to collecting benchmarking datasets and to make them useful to the larger community.

References

- 1 Dagstuhl seminar 19351 contributors. *Computational Proteomics*. <https://www.dagstuhl.de/19351>, Online; accessed 17-April-2020.

3.9 Mining Metabolome and Genome

Ewy Mathé (Ohio State University – Columbus, OH, US)

License  Creative Commons BY 3.0 Unported license
© Ewy Mathé

Many resources are available for supporting the integrated analysis of genomes and metabolomes. However, these resources are largely fragmented and mostly lack interoperability. Computational expertise is most often a requirement to piece together resources for appropriate interpretation of integrated metabolome-genome data. There is thus a need for defining common meta-data/controlled vocabulary, and for automating the process of deriving detailed meta-data for samples and analyte (metabolites, genes).

The task of this group was to define user cases and guidelines on how to use and integrate resources to meet user needs. Guidelines will include defining quantitative metrics to use these databases properly (e.g. FDR confidence, being able to detect discrepancies between different sources), and unit tests for data integration. The steps in integrating resources are modular. Limitations of each module were defined, so that users can then piece together different modules to meet their needs.

3.10 Confidence and Compound Identification


Hunter Moseley (University of Kentucky – Lexington, KY, US)

License  Creative Commons BY 3.0 Unported license
© Hunter Moseley

This session explored how to quantify confidence in compound identification. Three major types of confidence metrics were identified: confidence categories, (continuous) confidence scores, and probabilistic scores. Different types of probabilistic scores were covered, especially probabilistic scores that take into account false discovery. Inaccuracy in estimating low false discovery rates (FDR) in the context of MS/MS-based compound identification was discussed. An alternative or complementary method is to estimate compound identification ambiguity from assignment and dataset specific decoy generation. The supplementation of richer spectral data to improve assignment was mentioned. Using identification confidence and/or ambiguity to limit deposited annotations was discussed. The general consensus was that more assignment annotations with deposition would allow broader data reuse and improve interpretation.

3.11 MassBank: Status Meeting


Steffen Neumann (IPB – Halle, DE)

License  Creative Commons BY 3.0 Unported license
© Steffen Neumann

MassBank was the first open source, open data spectral library. Currently, there are sites in Japan, Europe, and the US (MoNA). In Dagstuhl there was the opportunity to discuss current and future developments among users, developers and related resources. These included the quality assurance in spectral library creation, an upcoming REST interface and opportunities to interchange data with other sites.

3.12 Autoencoders

Jamie Nunez (Pacific Northwest National Lab – Richland, WA, US) and Michael Andrej Stravs (Eawag – Dübendorf, CH)

License  Creative Commons BY 3.0 Unported license
© Jamie Nunez and Michael Andrej Stravs

Methods of generating potentially novel compounds was first covered, which included combinatorics, reactions, rule-based construction, experimental data-driven generation, and autoencoders. Autoencoders were covered in more detail, first describing their general set up and the use of latent space (a compressed version of the data input to the autoencoder which is then interpreted by the decoder). An example of structure-to-structure designs was examined, along with the considerations of its advantages and disadvantages, how training was done, and what latent space truly represents at the end. Other potential designs were then discussed, such as fingerprint-to-structure to generate candidates from experimental data and reactions-to-reactions. It is important to also keep in mind that decoders can often produce invalid output, which has to be checked, showing a need to carefully interpret the real meaning of the output and (non)continuity of latent space.

3.13 Collision Cross Section and Ion Mobility Spectrometry

Tomas Pluskal (Whitehead Institute – Cambridge, MA, US)

License  Creative Commons BY 3.0 Unported license
© Tomas Pluskal

Ion mobility spectrometry (IMS) is a technique for separating molecules in a neutral gas phase based on their drift time (time spent in the IMS chamber), which is proportional to the collision cross section (CCS) of the molecule. IMS can be conveniently combined with mass spectrometry for better separation and identification of molecules. Significant progress has been made in predicting CCS values of molecules using deep learning and quantum chemistry calculations. However, IMS presently suffers from relatively poor support in data processing tools and packages. During the session, various hardware approaches for IMS separation were introduced and specific needs for data processing tools were discussed. There was general consensus that IMS has great potential, but the current hardware and software capabilities are limited. In particular, a lack of a good algorithm for 4D (chromatography retention time,

IMS drift time, m/z , and intensity) feature detection was identified as a major bottleneck in the field. Development of new visualization tools and CSS distribution databases was also encouraged.

3.14 Jupyter Notebooks for #FAIR Data Science

Stacey N. Reinke (Edith Cowan University – Joondalup, AU)

License © Creative Commons BY 3.0 Unported license
© Stacey N. Reinke

The Jupyter Notebook is an open-source interactive coding tool that launches in a web browser. It contains cells for descriptive text and live code; outputs of executed code cells (tables, visualisations) are then displayed immediately below the code cell. This framework was developed to enable transparent sharing of code and workflows, therefore promoting FAIR data science in the scientific community. More recently, the launch of the Binder deployment service has allowed researchers to share their Jupyter Notebooks in the cloud with a url link. This session provided a description of Jupyter Notebooks and Binder, as well as their practical utility in workflow sharing and education.

3.15 Statistical Integration


Stacey N. Reinke (Edith Cowan University – Joondalup, AU)

License © Creative Commons BY 3.0 Unported license
© Stacey N. Reinke

Metabolomics data can be integrated with other types of data, such as other omics or clinical data, to enable a more comprehensive understanding of the biological system. This session aimed to identify and discuss different approaches for data integration of two or more matrices, one being metabolomics data. Three different approaches were identified. Network-driven integration approaches require *a priori* biological knowledge. They can include mathematical models of individual biological processes or pathway mapping. Pathway mapping often suffers from lack of interpretability due to the high level of metabolic interconnection. Dimension reduction integration aims to reduce the metabolic feature space prior to downstream pathway analysis; however, testing has shown lack of robustness with respect to pathway definition. Data-driven integration approaches include methods such as correlation and multivariate analyses. These approaches can enable the identification of novel biology; however, they are limited by lack of usability and interpretability. The outcome of this session included a list of tools for achieving data integration and also an acknowledgement that this is a developing field which needs to be further developed prior to large scale implementation.

3.16 Biosynthesis, Gene Clusters, and Predicting Natural Products from the Genome

Justin van der Hooft (Wageningen University, NL) and Simon Rogers (University of Glasgow, GB)

License  Creative Commons BY 3.0 Unported license
© Justin van der Hooft and Simon Rogers

Metabolite identification of natural products can be accelerated by linking information gained from genome sequences. This breakout group started with a short historical perspective on using structural information from the genome to inform structural elucidation which started back in 2005 with the first natural product being predicted from the genome of *Streptomyces coelicolor*. The major questions the group addressed were what structural and quantitative information can be predicted from genomes? And how do Biosynthetic Gene Clusters help? A list of resources included the PRISM and antiSMASH ecosystems that sparked the development of tools that link genome and metabolome data. Listed examples are GNP and NRPQuest and RiPPQuest that show relative successful examples for modular structures like peptides and some polyketide classes. The Dorrestein lab developed peptidogenomics and glycogenomics workflows that link the genome and metabolome by predicting amino acid and sugar moieties, respectively, that can be searched for in mass spectrometry data through mass differences and neutral losses. The group then discussed the next steps. Linking genomes directly to structures is a (very) hard problem; linking the genome to spectra is still challenging but can be regarded as “ranking problem”. In the genome, gene domains are mainly used to translate the genome into structural information – through (predicted) enzyme activity. In the metabolome/metabolomics, once annotated, structural elements (substructures) can be exploited, for example by using chemical fingerprints. However, spectral patterns on themselves could be used as well to link to specific genetic elements. This could be helpful in prioritizing candidate spectra – gene links found by correlation approaches (based on strain presence/absence). Finally, the group discussed how prioritization of candidate structures could be improved by allowing to select groups of metabolites from one organism or – more widely – from natural products – or even more generic – from molecules that could be found in nature – which could include pesticides. Altogether, accelerating natural product discovery through linking the genome to the metabolome is a promising field!

3.17 Bioschemas

Egon Willighagen (Maastricht University, NL)

License  Creative Commons BY 3.0 Unported license
© Egon Willighagen

Bioschemas (<https://bioschemas.org/>) is an extension of the schema.org standard used by major search engines like Google and Bing to recognize information or metadata they want to use in their indexing. Bioschemas has an annotation type for the life sciences, like Protein and MolecularEntity, but also types for Tool (e.g. software) and TrainingMaterial (like tutorials). It is supported by the EU ELIXIR community as an interoperability layer and is used in a variety of their projects. In this session we discussed what it is and is not (e.g. it is not an ontology), looked at various annotation types (called “profiles”), and what information one can add to it. We looked at various solutions people have found to use Bioschemas

in their project. For example, we looked at how Bioconductor package vignettes can be extended. Additionally, we looked at how ChEMBL uses Bioschemas on their HTML pages. The Bioschemas website has a page with live deployments. The meeting was concluded with hacking on Bioschemas annotation of Bioconductor packages itself, continuing a patch initiated at a computational metabolomics meeting in Wittenberg, DE in April 2019.

3.18 Open Biological Pathways with WikiPathways

Egon Willighagen (Maastricht University, NL)

License  Creative Commons BY 3.0 Unported license
© Egon Willighagen

WikiPathways (<https://wikipathways.org/>) is a free, online, community-driven, curated knowledgebase of biological pathways. Comparable and complementary to other databases like KEGG and Reactome, WikiPathways has a semantic representation of biological processes, resulting from past and current collaborations with research communities like WormBase, LIPIDMAPS, NetPath, EJP-RD, and many, many more. The WikiPathways Portals reflect this community embedding. The focus has always been on interoperability and semantic meaning which was discussed in the session. The semantic web format and SPARQL application programming interface were also discussed. We walked through a number of further integrations, such as EuropePMC linking to pathways for articles that are cited by that pathway (LabLinks), Wikidata, and named resources that include the WikiPathways, such as RAMP. Finally, we looked at how pathways are drawn with PathVisio, extended with CyTargetLinked in Cytoscape with transcription factors, miRNAs, and drugs (-lead) from DrugBank and ChEMBL. Questions around the underlying GPML format and RDF export were discussed, in addition to the curation process, and how all this is used in systems biological pathway and network enrichment analyses.

3.19 Retention Time

Michael Anton Witting (Helmholtz Zentrum – München, DE)

License  Creative Commons BY 3.0 Unported license
© Michael Anton Witting

Retention times represent an interesting orthogonal information for metabolite identification. However, they are less standardized compared to other parameters and represent a property of the metabolite and the employed chromatographic system in comparison to mass, which is a molecular property. In this session we discussed the current state of the art in retention time prediction and how it can be integrated with e.g. analysis of tandem MS data. An approach for prediction of retention orders developed by Juho Rouso and Sebastian Böcker was discussed and what kind of additional data is required to further develop it.

3.20 Conclusion: The Future of Computational Metabolomics

Sebastian Böcker (Friedrich-Schiller-Universität Jena, DE)

License  Creative Commons BY 3.0 Unported license
© Sebastian Böcker

In this plenary discussion, we tried to identify upcoming research questions in computational metabolomics, but also identify new possibilities that computational methods will provide for metabolomics in general. Computational methods for, say, small molecule annotation have evolved greatly in recent years, as demonstrated by CASMI contests [1]; how can we continue with method development in this speed, and how do we best utilize the developed methods? One particular topic of discussion was how to attract experts from machine learning to work on metabolomics problems. Here, it is of utmost importance to lower the barrier to enter the field for scientists from machine learning; e.g., to formalize problem(s) and to describe them in terms that machine learning scientists can understand (graph theory, optimization, etc). We will try to use the Kaggle platform (<https://www.kaggle.com/>) to attract ML experts; we identified some topics such as anomaly detection in clinical environments (i.e. high cholesterol) and retention time/order prediction as topics where this may be possible. Another topic of discussion was the disruptive changes of MS and computational technology: Where do we expect them to be, and what impact will these changes have? Discussed topics included prediction accuracy, quantum computing, the use of GPUs, and substantial increase in annotation rates. A third topic of discussion was metabolic modelling and stable isotope labelling experiments: these can lead to improved biological insight, with or without stable isotope labeling. We discussed the potential to use existing and new datasets that link metabolomics, transcript, genomics, or proteomics to improve interpretability; the importance of FAIR data was mentioned in this context. Finally, improved annotation can produce better metabolic/system modelling and allow us to generate new biological hypotheses.

References

- 1 Schymanski, Emma L., et al. Critical Assessment of Small Molecule Identification 2016: automated methods. *Journal of Cheminformatics*, 9.1 (2017): 22.

Participants

- Oliver Alka
Universität Tübingen, DE
- Nikiforos Alygizakis
Environmental Institute –
Koš, SK
- Sebastian Böcker
Universität Jena, DE
- Evan Bolton
National Institutes of Health –
Bethesda, US
- Corey Broeckling
Colorado State University –
Fort Collins, US
- Celine Brouard
INRA – Toulouse, FR
- Andrea Brunner
KWR Water Research Institute –
Nieuwegein, NL
- Jacques Corbeil
University Laval – Québec, CA
- Alexis Delabriere
ETH Zürich, CH
- Pieter Dorrestein
University of California –
San Diego, US
- Xiuxia Du
University of North Carolina –
Charlotte, US
- Timothy Ebbels
Imperial College London, GB
- Markus Fleischauer
Universität Jena, DE
- Laurent Gatto
University of Louvain, BE
- Kati Hanhineva
University of Kuopio, FI
- Rick Helmus
University of Amsterdam, NL
- Lukas Käll
KTH Royal Institute of
Technology – Solna, SE
- Oliver Kohlbacher
Universität Tübingen, DE
- Adelene Lai Shuen Lyn
University of Luxembourg, LU
- Jan Lisec
BAM – Berlin, DE
- Marcus Ludwig
Universität Jena, DE
- Tytus Mak
NIST – Gaithersburg, US
- Hiroshi Mamitsuka
Kyoto University, JP
- Ewy Mathé
Ohio State University –
Columbus, US
- Hunter Moseley
University of Kentucky –
Lexington, US
- Steffen Neumann
IPB – Halle, DE
- Louis-Felix Nothias-Scaglia
University of California –
San Diego, US
- Jamie Nunez
Pacific Northwest National Lab. –
Richland, US
- Tomas Pluskal
Whitehead Institute –
Cambridge, US
- Stacey N. Reinke
Edith Cowan University –
Joondalup, AU
- Simon Rogers
University of Glasgow, GB
- Juho Rousu
Aalto University, FI
- Augustin Scalbert
IARC – Lyon, FR
- Tobias Schulze
UFZ – Leipzig, DE
- Emma Schymanski
University of Luxembourg, LU
- Christoph Steinbeck
Universität Jena, DE
- Michael Andrej Stravs
Eawag – Dübendorf, CH
- Justin van der Hooff
Wageningen University, NL
- Philip Wenig
Lablicate – Hamburg, DE
- Egon Willighagen
Maastricht University, NL
- David Wishart
University of Alberta –
Edmonton, CA
- Michael Anton Witting
Helmholtz Zentrum –
München, DE
- Oscar Yanes
Rovira i Virgili University –
Reus, ES
- Nicola Zamboni
ETH Zürich, CH

