ISSN 2186-7437

NII Shonan Meeting Report

No. 2017-3

NII Shonan Meeting Report
Computational Metabolomics

Masanori Arita
Sebastian Böcker
Steffen Neumann

March 20–23, 2017

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

NII Shonan Meeting Report
Computational Metabolomics

Organizers (in alphabetical order):
Masanori Arita (National Institute of Genetics, Japan)
Sebastian Böcker (Friedrich Schiller University Jena, Germany)
Steffen Neumann (Leibniz Institute of Plant Biochemistry, Germany)

March 20–23, 2017

Overview

Metabolites are not only responsible for tasks such as growth, development, and reproduction, but also directly relevant to structure, signaling, and chemical interactions with other organisms. Metabolomics therefore plays an essential role in the omics sciences, investigation of novel drug leads or profiling metabolites of pharmaceutical compounds for their side effects. With advances in instrumentation, metabolomics is currently at the edge of becoming a "big data" science.

Mass spectrometry is the predominant analytical technique for detecting and identifying metabolites and other small molecules in highthroughput experiments. Huge technological advances in mass spectrometers and experimental workflows during the last decades enable novel investigations of biological systems on the metabolite level. These advances, however, also resulted in a tremendous increase of both amount and complexity of the experimental data. The data processing and metabolite identification form the largest bottlenecks in highthroughput metabolic analysis. Unlike proteomics, where close cooperations between experimental and computational scientists have been established over the last decade, such cooperation is still in its infancy for metabolomics. The key goal of this seminar was to foster the exchange of ideas between the experimental (analytical chemistry and biology) and computational (computer science and bioinformatics) communities.

The seminar was mostly organised as a single-track event, with several breakout groups. Important topics were solicited from all participants prior to the meeting and the importance of each topic was voted. According to the voting results, topics were arranged in the morning sessions and introduced by respective presenters (topic introduction). Then, details were fully discussed, based on the morning talks, in the afternoon.

The meeting outline was as follows.

- Day 0 evening
  - Welcome reception
- Day 1 morning
  - opening and overview
  - Topic introductions for database search, identification, and omics
- Day 1 afternoon
  - Breakout discussions for database and omics
  - Plenary discussion for compound identification
- Day 2 morning
  - Topic introductions for novel structures, retention time, statistics, and feature finding
- Day 2 afternoon
  - Plenary discussion for structure and retention
  - Plenary discussion for statistics and feature
- Day 3 morning
  - Topic introductions for deconvolution and standards
  - Plenary discussion
- Day 3 afternoon
  - Excursion to Kamakura
- Day 4 morning
  - Summary discussion and MoU signing

1 Talk abstracts / Position statements

Searching in Structure databases

Hiroshi Tsugawa (RIKEN Center for Sustainable Resource Science)
Juho Rousu (Aalto University)

Assigning reliable scores for ranking metabolites is important. Such scores are based on MS spectra, molecular structures, and the algorithm to connect them. Different software programs use different rankings and a criterion for reliable annotation is unknown. A simplistic approach will face difficulty in treating complex lipids, for example, the differentiation between phosphatidyl glycerol (PG) and bis(monoacylglycerol)phosphate (BMP). Good design of structure-based fingerprints is important, but the structure information only seems not enough. Since metabolites are identified in various samples from plants, animals and environments, one strategy is to exploit the information of bioactivity of metabolites. Such information can be exploited to at least exclude known metabolites of known activities.

Practical issues of compound ID

Steffen Neumann (IPB Halle)

MS/MS information offers an indispensable clue for identifying metabolites, but associated metadata, such as the acquisition method and strategy, offer clues to handle the spectral information. Such information should be integratively considered as metabolite ID for identification. Importance of associated information such as retention time, pathways and isotopic analyses will be discussed.

OMICS integration

Masanori Arita (NIG and RIKEN CSRS)

The field of lipidomics is rapidly expanding. The coordination with metabolomics field is strongly anticipated. Imminent issue is the shorthand name for lipid molecules. Lipidomics researchers tend to annotate stereochemically non-ambiguous names to their results, but such custom might conflict with the identification standard in the metabolomics community. Research collaboration and more discussion is necessary and only after such resolusion, we can proceed to integrate with transcriptomic and genomic data.

Meta-methods and tool integration

Sebastian Böcker (FSU Jena)

Meta methods are integrative approach of multiple tools and there are two directions: vertical and horizontal. In the vertical approach, software programs are pipelined to obtain results. Most important is the use of best available tool in each stage. In the horizontal approach, software programs for the same step are

multiplly run and compared. Since different tools show different performances, a rigorous cross validation is necessary to integrate horizontally parallel tools.

Computer-assisted structure elucidation (CASE)

Christoph Steinbeck (FSU Jena)

De-novo structure generation is important but full enumeration is combinatorially prohibitive and directed structure generation requires enough spectral information during the generation process. To overcome, several issues exist. First is the requirement of reaction-driven biospace expansion, as known from tools like BioTransformer (D. Wishart), MINEs, metaprint2d or the commercial Meteor and Derek software. In either case, CASE requires a rich set of Open Source cheminformatics tools and workflows. Useful tools include KNIME, RDKit and CDK.

FDR calculation for metabolomics

Sebastian Böcker (FSU Jena)

False Discovery Rate (FDR) estimation and related concepts (q-values, Posterior Error Probabilities) are still in their infancy in metabolomics; this is quite different than in proteomics, transcriptomics etc where these concepts have become integral parts of the everyday research. Novel method for FDR estimation when searching spectral libraries with tandem mass spectrometry data: Scheubert et al., bioRxiv 2017, doi 10.1101/109389.

Retention time in LC

Michael Witting (Anal BioGeoChem, HMGU)

Retention time information can be used for prioritization or rejection of candidate structure for in silico fragmentation. For a rough estimation, the octanol-water partition coefficient logP or logD can be used for predicting elution ranges, e.g. if metabolite will be retained under the given chromatographic conditions. The ultimate goal would be exact prediction of retention times. Different approaches have been used for different separation systems, e.g. by Creek et al., Cao et al. or Eugster et al. However, these systems only use small sets of metabolites ≥ 500 and have therefore only limited prediction capabilities. Additionally, prediction errors are quite high and usually do not allow the differentiation between closely eluting isomers (which often yield also similar fragmentation). Different possibilities to overcome current limitations are introduced.

Feature finding

Oliver Kohlbacher (Univ Tübingen), Tomáš Pluskal (MIT)

Obtaining all features that represent the same compound is crucial. In order to find consistent features, we can use a series of diluted samples to check their

stability. Use of mass difference of isotopes is another way. A standard format should include the following features: m/z and RT values, peak shape, charge state, isotopes, adducts, MS/MS spectra, fragmentation trees. Then we can start benchmarking datasets. Especially in MS/MS, statistics are necessary to evaluate the reliability of spectra.

Deconvolution

Hiroshi Tsugawa (RIKEN CSRS)

Deconvolution refers to a computational process to separate co-eluting chromatographic peaks to extract independent, original peaks. In MS-DIAL software, deconvolution of up to five co-eluting peaks is available for data from not only GC but LC/MS studies. The minimum distance between separable peaks are two scans. The basic idea and the optimization method of the deconvolution are introduced.

Data standards

Oliver Kohlbacher (Univ Tübingen) and Reza Salek (EBI)

The PhenoMeNal (Phenome and Metabolome aNalysis) H2020 e-infrastructure project and the Netherlands Metabolomics Centre, in coordination with ELIXIR-NL and ELIXIR-DE, are jointly organising a strategic "Workshop On Establishing a Metabolomics Use Case in ELIXIR" on the 25th April 2017 in Frankfurt, Germany. In the context of ELIXIR, the main goal of the workshop is to identify the main needs of the metabolomics community (at different levels) in the context of a life science computing related infrastructure.

2 Summary of discussions

Searching database

Identification scores produced by software tools depend on spectral data quality, compound class, and algorithm. An informative MS/MS spectrum contains distinct features both 'specific' and 'generic'. A good spectrum contains peaks across the entire spectrum, and the database search should exploit them all. Adduct information, measurement conditions (pH and so on), and separation chemistry are all important in the search process.

OMICS integration

Generalization of concepts across omics is necessary, and concepts need names. There are range of problems: the baseline issue is how to accurately link different entities via standardised names. More specifically, the lack of datasets that have adequate temporal resolution to track changes between "levels" complicates the linking problem. Metabolite concentrations themselves are not very informative, we need to measure fluxes within pathways.

One exemplary website for data integration is the Cancer Genome Atlas for predicting phenotypes from omics data. Omics Discovery Index (Omics DI; http://www.omicsdi.org) can use such resources (metadata only) for locating related clues.

Compound identification

Choosing the right molecular formula is a key issue in identification. For the minimum workflow for identification, there are several examples.

- Example in Glasgow
- Phenome center UK: standardized protocol produces a report
- MetaboLight automated identification toolkit (Trinity tool Galaxy plug)

GC/MS is fully automated already, and automation is hard for unknown metabolites. For the efficiency of software tools, benchmarking is important and the size of databases matters. A few thousand dataset is too small.

Substructure space

In natural product research, the way that the community looks at substructure space may be biased. We need tools for structure generation that can produce realistic candidates and they should exploit existing tools and databases such as BioTransformer, MINE, KNIME, and FooDB.

Representation of uncertain structure

It is simple to report a particular metabolite structure, either by using an accession in one of the established molecular structure databases, or as the structure itself using a SMILES string or preferably the IUPAC recommended InChI. These representations can include or exclude the specific stereo configuration of the molecule, but they do not allow to report uncertain structures.

A common case is that several positional isomers are difficult or even impossible to resolve with the given spectral information. In such cases, chemists use a notation to place a chemical group at an arbitrary place in the ring structure (Markush structure) but this scheme is not machine-readable. Support for such abstract notation is necessary (https://www.chemaxon.com/library/ markush-structures- at-chemaxon/). If there are few isomers, a simple enumeration, with e.g. SMILES, is possible.

Retention time information in LC/MS

During the meeting, different possibilities to overcome current limitations have been discussed. First, better sharing and reporting of retention time data would increase its reuse. Second, the integration of different data sets into machine learning approaches was discussed to overcome the limited number of metabolites and improve predictive power.

Additional information commonly used by experts is the ionisation behaviour of different compound classes, especially when it is possible to obtain spectra

with both ESI and APCI ionisation, or whether some adduct forms ($[M+H]^+$, $[M+Na]^+$ or $[M+NH4]^+$) are more likely given the compound class and experimental condition. Predicting these will require new cheminformatics tools to predict ionisation behaviour.

FDR calculation for metabolomics

False Discovery Rate (FDR) estimation and related concepts (q-values, Posterior Error Probabilities) are still in their infancy in metabolomics; this is quite different than in proteomics, transcriptomics etc where these concepts have become integral parts of the everyday research. Sebastian Böcker gave a general introduction to these concepts, and also presented a novel method for FDR estimation when searching spectral libraries with tandem mass spectrometry data (Scheubert et al., bioRxiv 2017, doi 10.1101/109389).

Feature detection / extraction

There is no standardised format for feature detection output. For benchmarks, simulated LC/MS data are useful. We also require re-creation of data for new instrumental characteristics. It is somewhat similar to the "Turing test" for data: calculation of statistical characteristics to determine how "real" they are. Such decoy construction is crucial in many occasions.

Database curation

Many databases are missing feedback channels. There are some simple sanity checks / validation possible, like molecular mass matches molecular formula matches molecular structure. It is better to associate such software development practices into databases. For example, use unit test like checks, MoNA is doing dozens of (simple) tests. Use version control systems for change management. Have clearly defined interfaces for automated validators of individual aspects, maybe have hackathons / jamborees to add validators to the databases.

Statistical analysis

Analysis is complicated and needs a bridge among people involved. Consider the problem of missing values / imputation methods. Various options are available for dealing with missing values such as imputation or using test statistics. However, these methods are generally conceptually advanced, complex to apply and inteprete and this mitigates against their use by those who are not familiar with. Conversely, many statisticians find these problem uninteresting or trivial. A related issue is raising "statistical consciousness" and "know how" among those who will need to perform these analyses. A good place to start is identification of some key problems that couod form the basis for an informative "open problem" type article.

Additional topics in Day 4

Among the main outcomes are the creation of initiative to organise and collect several contributions to the Encyclopedia of Lipids. In addition, the 24 attendees signed a Memorandum of Understanding, with the following goals:

- initiating a long-term research cooperation to advance metabolomics science and to the benefit of the community at large
- cooperation in the field of Computational Metabolomics and Mass Spectrometry
- promotion of researchers and students exchange, open data standards, and open data sharing; organization of joint symposia at least once every three years on the above topics.

Appendix: List of participants

- Rainer Breitling (University of Manchester)
- Celine Brouard (Aalto University, Espoo)
- Kai Dührkop (FSU Jena)
- Timothy Ebbels (Imperial College, London)
- Tobias Kind (UC Davis)
- Oliver Kohlbacher (University of Tübingen)
- Kris Morreel (Ghent University)
- Tomas Pluskal (MIT, Boston)
- Juho Rousu (Aalto University, Espoo)
- Nozomu Sakurai (Kazusa DNA)
- Reza Salek (EBI)
- Emma Schymanski (Eawag, Duebendorf)
- Christoph Steinbeck (EMBL-EBI Cambridge)
- Hiroshi Tsugawa (RIKEN)
- Michael Witting (Helmholtz Zentrum M?nchen)
- Akiyasu Yoshizawa (Kyoto U)
- Satoshi Tanaka (Trans-IT Inc.)
- Justin van der Hooft (University of Glasgow)
- Bo Johannes BURLA (National University of Singapore)
- Jacques Corbeil (University Lavel, Quebec)
- Rohan B.H. Williams (SCELSE Singapore)
- Shin Kawano (DBCLS, ROIS)
- Atsushi Fukushima (RIKEN CSRS)

- Kati Hanhineva (University of Eastern Finland)
- Ville Koistinen (University of Eastern Finland)
- Masanori Arita (National Institute of Genetics)
- Sebastian Böcker (Friedrich Schiller University Jena)
- Steffen Neumann (Leibniz Institute of Plant Biochemistry, IPB Halle)