



Bioinformatics can boost metabolomics research



René Meier^a, Christoph Ruttkies^a, Hendrik Treutler^a, Steffen Neumann^{a,b,*}

^a Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany

^b German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig Deutscher, Platz 5e, 04103 Leipzig, Germany

ARTICLE INFO

Keywords:

Metabolomics
Mass spectrometry
Metabolite profiling
Metabolite identification
Workflows
Cloud computing

ABSTRACT

Metabolomics is the modern term for the field of small molecule research in biology and biochemistry. Currently, metabolomics is undergoing a transition where the classic analytical chemistry is combined with modern cheminformatics and bioinformatics methods, paving the way for large-scale data analysis. We give some background on past developments, highlight current state-of-the-art approaches, and give a perspective on future requirements.

1. Introduction

Research in the modern life-sciences aims at the understanding of living organisms, where all processes between the genome and the phenotype are of interest. The subject of studies include gene regulation, protein synthesis, their post-translational modifications, and the biochemistry of proteins and small molecules – metabolites.

Metabolomics is the modern term for the field of small molecule analytical chemistry in biology and biomedical research, but the underlying questions have been addressed already for hundreds of years by physicians using the smell and colour of, e.g. urine for health diagnosis. In 1971, Pauling et al. (1971) analysed more than 200 metabolites in breath and urine headspace, but the terms “metabolomics” or “metabonomics” only appeared in the scientific literature more than 25 years later (Oliver et al., 1998; Nicholson et al., 1999). In the last two decades, a lot of progress has been made in the number of metabolites that can be (simultaneously) detected, lowering of the limits of detection with modern analytical technologies, and the increased throughput of samples that can be processed.

2. Scaling up mass spectrometry based metabolomics

Today, mass spectrometry is a key technology for metabolomics research. Due to immense technological advances in mass spectrometry over the last years, the amount and complexity of the data produced has been growing rapidly. These advances would not have been possible without the extensive use of computers throughout the data processing and analysis steps of the experiments.

While the first mass spectrometers used photo platters to record

spectra, the computer became an integral part of the instruments under the term “data system” already in the 1970s. The rapid digital recording of mass spectra also allowed to couple the MS instruments to chromatographic separation, such as liquid or gas chromatography. These separation processes greatly reduce the complexity of the individual mass spectra, which in turn allows to measure more complex samples, such as full methanolic extracts of plants or human body fluids. The advent of ion mobility further increases the separation power, but also increases the run time per sample.

The amount of data from raw spectra in LC/MS measurements is overwhelming, hence a feature detection step is typically applied to extract the chromatographic and spectral peaks into so called feature lists. These feature lists can be used as metabolic fingerprints, which represent a molecular phenotype. Typical metabolomics experimental designs include the comparison of different genotypes, perform intervention studies and time-series experiments. These setups require the processing of dozens to hundreds, even thousands of samples. With microarrays it is possible to quantify the abundance of RNA of specified (short) sequences and directly compare the gene expression across samples. In LC/MS and GC/MS however, the feature lists need to be matched across samples, and both chromatographic shifts and mass deviations have to be considered or even compensated for. Several academic software packages have been developed in the past years, with MetAlign being one of the first tools, initially developed for GC/MS processing, and later also adapted for high-resolution LC/MS data (Tikunov et al., 2005; Lommen and Kools, 2012). Now, both the XCMS package (Smith et al., 2006), developed as part of the Bioconductor project (Gentleman et al., 2003), and the OpenMS (Sturm et al., 2008; Röst et al., 2016) framework are celebrating their tenth anniversary.

* Corresponding author.

E-mail address: sneumann@ipb-halle.de (S. Neumann).

<http://dx.doi.org/10.1016/j.jbiotec.2017.05.018>

Received 18 February 2017; Received in revised form 10 May 2017; Accepted 16 May 2017

Available online 26 May 2017

0168-1656/ © 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

While previously experimental analysis comprised comparing just, say, two measurements, these new paradigms automate the data processing steps and allow the use of statistical analysis, giving confidence values for the discovery and interpretation of, e.g. biomarkers. An example is the study conducted by Thévenot et al. (2015), where more than 200 samples were processed with an advanced workflow comprising preprocessing, signal drift correction and batch effect removal, and uni- and multivariate statistics. In addition to the huge number of R packages available for statistics and modelling, user-friendly and advanced data analysis tools exist in the form of, e.g. MetaboAnalyst (Xia et al., 2015) or MetExplore (Cottret et al., 2010).

The feature detection step is followed by the annotation of the molecular structures, which are required for the biochemical interpretation. A main advantage here is that mass spectrometry is independent of the availability of the genome sequence, and can be applied to any organism and tissue type. On the other hand, both analytical limitations and the chemical diversity of metabolites, biochemical processes, and external influences prevent that all possible molecular structures present in an organism are known *a priori*.

Thus, metabolite identification is an important task in computational metabolomics. For several organisms, including human and model organisms such as *Arabidopsis thaliana*, metabolite databases have been established (Wishart et al., 2013; Mueller et al., 2003). If the compound is assumed to be known in these databases, it will be returned with a rather simple search for metabolites having a mass within an instrument-dependent error window. However, all compounds with a similar mass (and of course all with the same molecular formula) will be retrieved as false positive hits. Their number can be reduced if the molecular formula itself can be deduced from the accurate mass, isotopic pattern and further hints. Generic chemical databases like PubChem or ChemSpider (Wang et al., 2009; Pence and Williams, 2010) contain orders of magnitudes more structures, but also contain many, if not the majority, of compounds usually not relevant to the experimental question.

3. Metabolite identification

Complementing to GC/MS and LC/MS, more structural information is available from higher-order mass spectra, such as tandem MS or MSⁿ. Here the analyte ions undergo fragmentation in the instrument, and the fragmentation spectra provide a fingerprint of the molecular structure. Those spectra can be compared against reference data to identify the metabolite, but require that the compounds are available and reference data have been deposited in, e.g., MassBank (Horai et al., 2010), HMDB (Wishart et al., 2007), and METLIN (Smith et al., 2005), or one of the commercial offerings. An overview of spectral libraries and comparison of their chemical coverage can be found in Vinaixa et al. (2016).

While mass spectral libraries are growing for tandem MS or MSⁿ spectra, the coverage is still relatively small compared to the number of compounds that could potentially be present in typical samples. Especially if no reference data are available, the spectra previously have been interpreted manually, and structural hints constrained the set of possible molecular structures.

One of the aims in computational mass spectrometry is to fill this gap by proposing tentative identifications for unknown compounds. Different approaches exist for the identification process. Besides the prediction of molecular properties from the spectral information by rule based expert knowledge or by machine learning approaches and their matching with candidates from compound databases it is possible to reproduce the process of fragmentation *in silico*.

MetFrag, launched in 2010, was one of the first approaches to address this problem for accurate tandem mass spectra for hundreds of candidate structures from chemical databases (Wolf et al., 2010). MetFusion combines *in silico* fragmentation with spectral similarity search in MassBank (Horai et al., 2010) and has shown to be an excellent way to benefit from two different information sources.

Compound identification is a time-consuming task as it requires the look-up and combination of many different information sources to collect as many evidences as possible to support a putative identification. Especially for high-throughput analysis where hundreds of metabolites need to be identified the workload for analysts is notably high. Therefore, MetFrag was further enhanced to provide a methodological interface to query different databases and combine the information drawn into the identification process. These new functions greatly reduce the burden on users to collect and merge ever increasing amounts of information available for substances present in different compound databases, thus enabling them to consider more evidence. With parallelizing the MetFrag analysis the processing of hundreds of tandem MS spectra is performed within minutes and the addition of different information types has shown to improve identification rates from 6% up to 70% depending on the dataset and information sources used (Ruttkies et al., 2016).

The usage of computational approaches for the identification of metabolites has been shown in several studies (McMillan et al., 2016; Van Meulebroek et al., 2016; Narduzzi et al., 2015). So far, results from *in silico* methods alone are not sufficient to count as full identification and additional validation in the lab, preferably with an authentic standard, is required. However, using these tools augmented with the experimental context and additional meta data reduces the workload by providing high quality putative metabolite annotations.

With untargeted MS measurements from modern MS instruments it is possible to measure thousands of features along with their fragment spectra. Hence, the manual inspection and identification of these data sets is no longer an option. Recent developments aim at relating hundreds or thousands of unidentified features by spectral similarity. This approach results in clusters of biochemically related metabolites offering a bird's eye view on all feature classes in a sample. Molecular networking implements this idea by using a graph with features as nodes which are connected by edges if the spectral similarity is above a certain threshold (Watrous et al., 2012). This allows the quick inspection of classes of metabolites with many members and has been demonstrated on data from different organisms. MS2LDA follows a different strategy by identifying patterns in fragment spectra as fingerprints of chemical substructures (van der Hooft et al., 2016). This allows the decomposition of fragment spectra and the assignment of fragments to certain parts of the molecule. MetFamily performs a hierarchical clustering of features based on spectral similarity combined with principal component analysis of MS abundances (Treutler et al., 2016). This allows the discovery of biochemically related features with regulated behaviour under different conditions called regulated metabolite families.

4. Reproducible and shareable research

The scientific discourse through letters among researchers and articles in scientific journals has a long history going back centuries, but electronic data publications have emerged only in the last few years. The amounts of data recorded in the life sciences mandate that data is available and enriched with experimental metadata. In metabolomics, public repositories are available for several years now: The NIH funded Metabolomics Workbench (Sud et al., 2016) and the European MetaboLights (Haug et al., 2013) hosted at EMBL-EBI.

Even though various analysis tools exist to tackle these problems a standardization of the computational workflows is hardly implemented. The community's need for standardization becomes even more obvious when realizing that even today the reproducibility of results available in peer-reviewed publications is not always possible. Besides the experimental conditions, making computational analysis pipelines reproducible with standardization is a first step to solve this issue (Sandve et al., 2013).

Especially in metabolomics, the large number of samples and features in the experimental results rises the need for unattended data

Download English Version:

<https://daneshyari.com/en/article/6451851>

Download Persian Version:

<https://daneshyari.com/article/6451851>

[Daneshyari.com](https://daneshyari.com)