MetHouse: Raw and Preprocessed Mass Spectrometry Data

Antje Gaida, Steffen Neumann

Leibniz Institute of Plant Biochemistry,
Department of Stress and Developmental Biology
[agaida|sneumann]@IPB-Halle.DE

Abstract

We are developing a vendor-independent archive and on top of that a data warehouse for mass spectrometry metabolomics data. The archive schema resembles the community-developed object model, the Java implementation of the model classes, and an editor (for both mzData XML files and the database) have been generated using the Eclipse Modeling Framework. Persistence is handled by the JDO2 -compliant framework JPOX. The main content of the Data Warehouse are the results of the signal processing and peak-picking tasks, carried out using the XCMS package from Bioconductor, putative identification and mass decomposition are added to the warehouse afterwards.

We present the system architecture, current content, performance observations and describe the analysis tools on top of the warehouse.

Availability: `http://msbi.ipb-halle.de/`

1 Introduction

Mass spectrometry (both GC-MS and LC-MS) has become the workhorse technology for Metabolomics, measuring the abundance of a large number of metabolites in parallel.

For hypothesis-driven and targeted experiments spreadsheets and other "light-weight" storage and processing mechanisms are usually sufficient. With the ambitious goal of metabolomics covering the whole range of metabolites and quest for data-driven analysis, structured and high-performance data storage is mandatory.

Recent developments in the metabolomics community have led to data exchange standards like mzData [10] and mzXML [11], which are currently being merged. Several databases have been created for mass spectra, such as the METLIN database [14] for FTICR and MS^2 spectra, or the BinBase and SetupX system [2] for GC-MS data.

Datawarehouses are used to integrate data from multiple sources and operative (OLTP) database systems optimized for retrieval and analysis (OLAP). The BioMart system [7] provides a framework for building and querying large biological databases, with both a web-frontend, standalone design- and querying tools and a flexible command line interpreter. The BioMart was initially designed for the EnsEMBL sequence repository. For "green bioinformatics" the Plant Data Warehouse (PDW) has been created [4], which covers data on plant phenotypes, sequences and expression levels.

This paper is structured as follows: in the next section we give an overview of the kinds of metabolomics data we incorporate into the MetHouse system, followed by the description of the data preparation and import steps. We finish with a conclusion and outlook.

2 Metabolomics Data

A typical metabolomics experiment measures and compares multiple samples on a GC-MS or LC-MS machine. The raw signals have to be processed and aligned to be comparable across the runs. For a successful biological interpretation the signals have to be annotated and if possible identified.

2.1 Experimental Metadata

In a classic, hypothesis driven experiment, several plants are grown and the extracts are measured for their metabolite content. To be useful beyond this individual experiment, the MS data needs to be annotated with the experimental metadata. We have chosen the *Ar*chitecture for *Met*abolomics (ArMet) model [6] to capture the biological source of the subjects under study, the growth condition and treatment history and follows all steps of the samples towards the machine analysis. Furthermore each piece of information is connected to the large scale experiment, the person responsible for the experimental step and a time stamp. A database-enabled infrastructure is described in [8].

In a data-driven scenario, all previously recorded experiments can be mined for repeating patterns and high correlation between the biological context and the measurements. Those findings can be used as additional functional annotation which was not in the focus of the original experiments.

Capturing the history of the experiment is also crucial for quality control purposes. If, for example, a set of plants is clearly different regardless of the respective treatment, correlation with the experimental metadata could reveal two sets of seed batches being used. Or in a medical setting where the response of guinea pigs to the therapy is observed, unexpected metabolic states could be connected with a special diet (“treats”) on Thursdays.

2.2 Mass Spectrometry Data

In both GC-MS and LC-MS the samples are first separated by a chromatographic column. The duration of a typical experiment can vary between a few minutes and an hour. The chromatographic column is connected to the mass spectrometer, and mass spectra are taken at frequencies typically between 100s^{-1} and 0.5s^{-1}.

We use the term raw data for the mass spectrometry full scan data, as exported from the machine and shown in figure 1. This can be exported from the machine, and converted into the data exchange format mzData. The conversion adds meta-data to the spectra, such as the Spectrum Type, MS-level and Polarity.

However, the raw data is too detailed for biological interpretation, where only the integrated intensities are of interest. Therefore the raw data is subjected to a set signal processing steps, which extract a baseline, filter noise, detect and quantify individual peaks. Often this step is time-consuming, and hence the peak data needs to be stored along with the raw data and the software parameters used to create them.

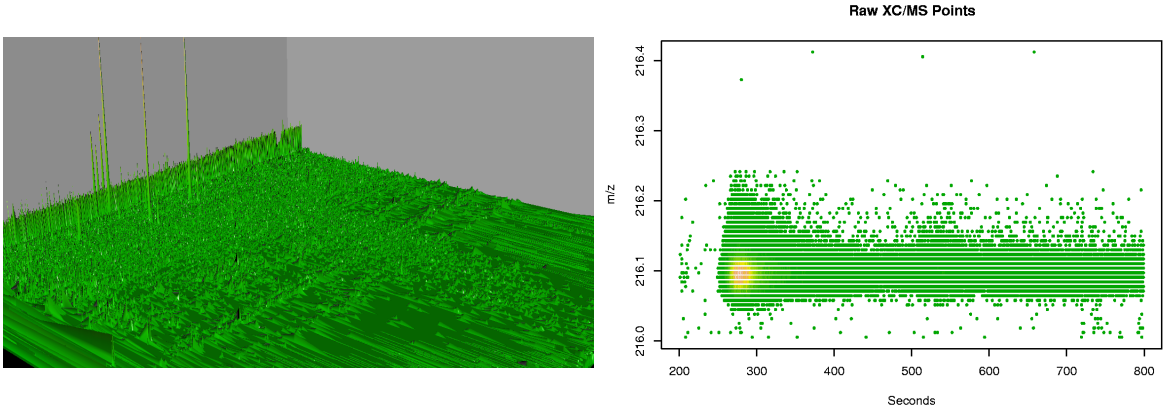

Figure 1: Left: 3D representation of LC-MS run typically with 4.5 to 5 million data points. The XCMS package condenses this into few thousand peaks. Right: Close-up of raw data for an arbitrary peak with mass mz=216.1 at time RT=290 sec.

2.3 Metabolite Identification

The identification of metabolites can be performed with different and complementary approaches.

Where libraries of spectra from identified compounds exist, a database lookup based on the mass spectra can be performed. For GC-MS based metabolomics there is a solid and growing set of compounds in e.g. the commercially available NIST library[1] or plant-specific selections in the Golm Metabolome Database[2] (GMD) [9]. The KNApSAcK system [12] includes chemical information on a large number of secondary metabolites, and also an online- or standalone browser. For LC-ESI-MS those libraries are often collected in-house.

For high-resolution mass data *de-novo* identification is possible, because the exact mass is a linear combination of the individual atom masses for each element. In addition, the isotope distribution can be taken into account to filter the initial result set. Such a package is DISOP (Decomposition of ISOtope Patterns) by Böcker et al. [1]. Others are included e.g. in vendor specific software, usually not easy to integrate in a software pipeline.

3 Implementation

In this section we describe the modules which together comprise the MetHouse. This includes archival of the raw data as exported from the mass spectrometer, and import of processed peaks into the warehouse. Finally, the peak data has to be connected to annotation, including links to chemical databases for identified compounds. An overview of the architecture is shown in figure 2.

[1] `http://www.nist.gov/srd/nist1.htm`
[2] `http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html`

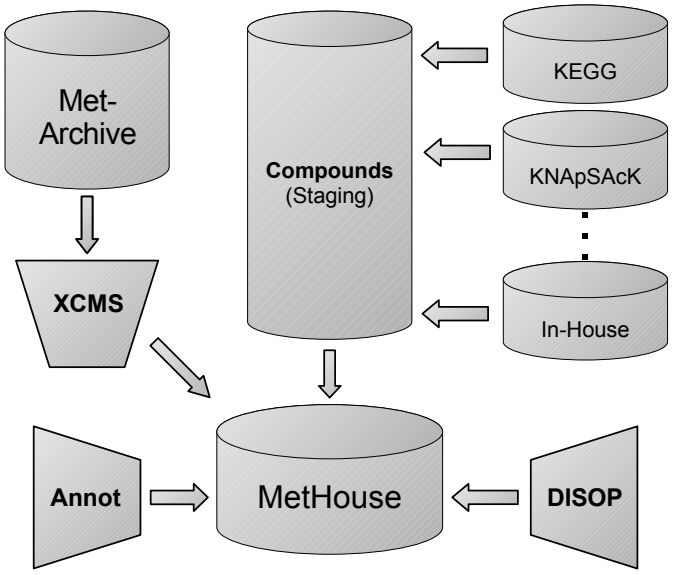

Figure 2: Overview of the MetHouse architecture. The output of the processing modules XCMS, DISOP, and Annot are imported into the datawarehouse. Multiple sources for Metabolite Identification are extracted from the original sources, transformed into a common schema and loaded into the datawarehouse.

3.1 Peak storage and processing

For a detailed description of our raw data archive see [8]. It is designed to store the mass spectrometry raw data, including both complete LC-MS runs and individual MS^n spectra taken from peaks of interest. Adhering to the mzData standard guarantees both the availability of converters (either from machine vendors or third parties) and at the same time detailed metadata (e.g. machine parameters) in the file. We also performed a benchmarking to show the timing for both import and retrieval of raw data, shown in figure 3.

Signal processing is part of the preprocessing step that operates on the raw data. The signal processing is done using the Bioconductor package XCMS [13].

A peak is described by its centroid expressed in mass and retention time, the minimum/maximum for mass and retention time and its intensities (using several quantification methods). The XCMS-parameter settings are stored along with the peak data, so peaks from a single input can be processed multiple times and distinguished or even filtered based on those settings.

Peak data is imported from within the R environment, so no temporary files have to be cre-

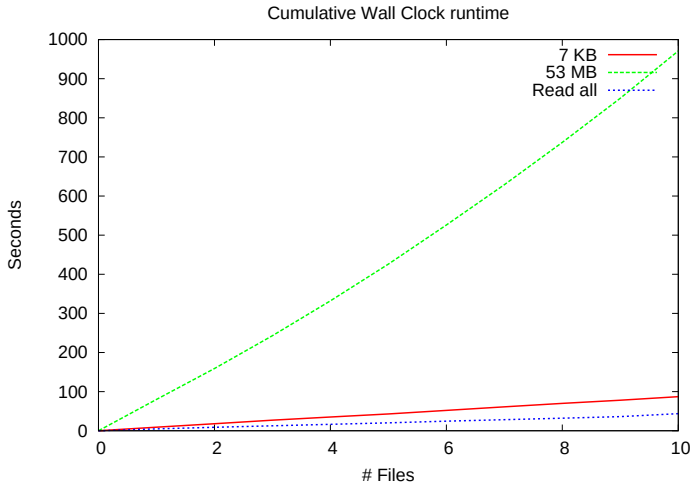

Figure 3: DB and import application. The small test file has 74 lines of XML using 7KB (1 Scan with 173 Peaks), the large file 36.000 lines in 53MB (1800 Scans with 5 Mio Peaks). Indices and constraints not disabled during import.

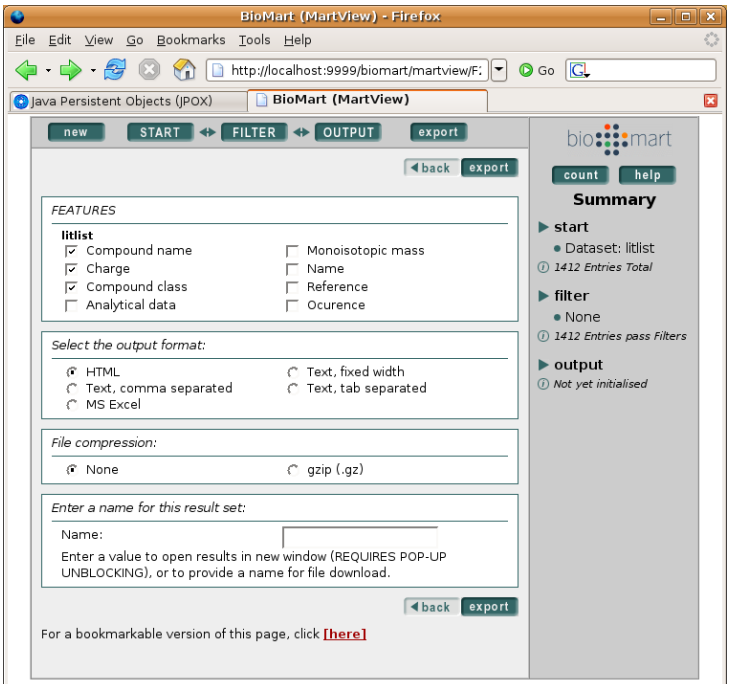

Figure 4: BioMart frontend for retrieval of compound information. The *litlist* has information on arabidopsis secondary metabolites compiled from the literature, and includes modifications expected in ESI-LC-MS.

ated. Since the database is connected via RODBC, many different RDBMS can be used for storage. For Oracle servers the native database interface (ROracle) has to be used, unless the commercially available ODBC drivers are installed. To improve speed, the import is done using prepared statements. However, usually the signal processing steps are the limiting factor.

3.2 Identification

Compound libraries are imported by parsing the respective flat files and kept in the staging area. For libraries in the NIST format *.MSP this is done using perl and the perl-DBI database API. The in-house library of MS^2 spectra is already kept in an RDBMS.

Before transfer into the data warehouse, the records are filtered (e.g. compounds irrelevant to plant research) or corrected/augmented (e.g. recalculating the exact mass in KEGG, which provides only two decimals for the mass). The isotope pattern can also be added, based on the elemental composition of the molecule.

The DISOP library is written in optimized C++ code. We wrapped the DISOP library into an R-package, so it can easily operate on both the XCMS output and MetHouse database content.

3.3 Data Warehouse

The data warehouse is built around the BioMart, which is a powerful and open-source framework for biological datawarehouses, and integrates well with both the Bioconductor project [3] and the Taverna workflow system [5]. A screenshot of the web interface is shown in figure 4.

Possible queries on the peaks are e.g. the retrieval of mass and retention time of the internal standards, to check for the stability of the machine calibration. Currently the data warehouse contains the data shown in table 1. The server is an AMD64 X2, 2GB Ram, 4*300GB SATA disks in RAID5 configuration running PostgreSQL 8.1.4.

Objects	Content	Source
Compounds LitList	1412	In-House Excel Sheet & Macro
Compounds MS^2	89	In-House Spectra Library
Compounds GMD	1166	`csbdb.mpimp-golm.mpg.de`
Compounds KEGG	1100	`www.genome.jp`
Compounds KNApSAcK	99273	`kanaya.aist-nara.ac.jp`
LC-MS Experiments	41	MzData Archive
Processed Peaks	52778	XCMS peak picking

Table 1: Current number of objects in MetHouse. The compounds already imported have been selected for their relevance to arabidopsis metabolomics research. The number of peaks per experiment ranges from 1203 to 1426.

4 Summary & Outlook

We have chosen and combined several open and extensible technologies for RDBMS, ORM, signal processing, data warehousing and statistics. The MetHouse system is currently under development, and even in this early stage includes a number of sources for each of the required steps in metabolomics data analysis.

It is also possible to store peaks for which no raw data is available, i.e. to circumvent XCMS preprocessing. In this case the peak lists have to be transformed during the ETL process to conform with the schema. Some post-processing, e.g. chromatogram based peak-shape correlation calculations are then not possible.

The planned inclusion of e.g. direct-injection FTICR[3] will also (by definition) have to replace XCMS preprocessing, but will equally benefit from the compound identification.

Further work will improve the connectivity to efforts currently developed in the metabolomics community, such as BioMoby services for the individual data sources, or processing services for mass spectrometry data. This way, complete mass spectrometry workflows can be created, using e.g. the GUI from the Taverna project. Eventually, joining data from the available -omics technologies will put functional genomics on the fast track.

Acknowledgements

Thanks to Nigel Hardy, Helen Jenkins, Chris Taylor, Kai Runte and many others for the ArMet and MzData models and Dierk Scheel, Jürgen Schmidt for their valuable discussions.

The work is supported under BMBF grant 0312706G.

References

[1] S. Böcker, M. Letzel, Zs. Lipták, and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proceedings of the 6th Workshop on Algorithms in Bioinformatics WABI 2006*,

[3]Fourier Transform Ion Cyclotron Resonance Mass Spectrometry, an (expensive) technology to measure masses at a very high resolution.

volume 4175 of *LNBI*. Springer, 2006. To appear.

[2] S. M. Fiehn O, Wohlgemuth G. Automatic annotation of metabolomic mass spectra by integrating experimental metadata. In *Proceedings of DILS 2005*, number 3615 in Proc. Lect. Notes Bioinformatics, pages 224–239. Springer, 2005.

[3] R. C. Gentleman, V. J. Carey, D. M. B., B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[4] I. Grosse, T. Funke, C. Kuenne, S. Neumann, A. Stephanik, T. Thiel, and S. Weise. Integrative Datenanalyse mit dem Plant Data Warehouse. *Vorträge für Pflanzenzüchtung*, 70:50–53, 2006.

[5] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucl. Acids Res.*, 34(suppl. 2):W729–732, 2006.

[6] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, and D. B. Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*, 22(12):1601–1606, December 2004.

[7] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Res.*, 14(1):160–169, 2004.

[8] S. Klie and S. Neumann. Storage and processing of mass spectrometry data. In *Proc. of 17th Int. Conference on Databases and Expert Systems (DEXA 2006)*, pages 211–215. DEXA, IEEE, September 2006.

[9] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, 21(8):1635–1638, 2005.

[10] S. Orchard, H. Hermjakob, P. Binz, C. Hoogland, C. Taylor, W. Zhu, R. J. Julian, and R. Apweiler. Further steps towards data standardisation. *Proteomics*, 5(2):337–339, 2005.

[11] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*, 22(11):1459–66, 2004.

[12] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya. *Plant Metabolomics*, chapter KNApSAcK: A comprehensive species-metabolite relationship database., pages 165–181. Biotechnology in Agriculture and Forestry. Springer, 2006.

[13] C. Smith, E. Want, G. O'Maille, R. Abagyan, and G. Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry*, 78(3):779–787, 2006.

[14] C. A. Smith, G. O. Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. Metlin: A metabolite mass spectral database. In *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring & Clinical Toxicology*, volume 27, pages 747–751, Louisville, Kentucky, April 2005.