MeltDB: a software platform for the analysis and integration of metabolomics experiment data

Heiko Neuweger1,∗, Stefan P. Albaum2, Michael Dondrup2, Marcus Persicke5, Tony Watt1,3, Karsten Niehaus3, Jens Stoye4 and Alexander Goesmann2
1International NRW Graduate School in Bioinformatics and Genome Research, Bielefeld University, 2Computational Genomics, Center for Biotechnology (CeBiTec), 3Proteome and Metabolome Research, 4Genome Informatics, Bielefeld University, and 5Institute for Genome Research and Systems Biology, CeBiTec, D-33501 Bielefeld, Germany

Received on June 23, 2008; revised on August 10, 2008; accepted on August 19, 2008
Advance Access publication September 2, 2008
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The recent advances in metabolomics have created the potential to measure the levels of hundreds of metabolites which are the end products of cellular regulatory processes. The automation of the sample acquisition and subsequent analysis in high-throughput instruments that are capable of measuring metabolites is posing a challenge on the necessary systematic storage and computational processing of the experimental datasets. Whereas a multitude of specialized software systems for individual instruments and preprocessing methods exists, there is clearly a need for a free and platform-independent system that allows the standardized and integrated storage and analysis of data obtained from metabolomics experiments. Currently there exists no such system that on the one hand supports preprocessing of raw datasets but also allows to visualize and integrate the results of higher level statistical analyses within a functional genomics context.

Results: To facilitate the systematic storage, analysis and integration of metabolomics experiments, we have implemented MeltDB, a web-based software platform for the analysis and annotation of datasets from metabolomics experiments. MeltDB supports open file formats (netCDF, mzXML, mzDATA) and facilitates the integration and evaluation of existing preprocessing methods. The system provides researchers with means to consistently describe and store their experimental datasets. Comprehensive analysis and visualization features of metabolomics datasets are offered to the community through a web-based user interface. The system covers the process from raw data to the visualization of results in a knowledge-based background and is integrated into the context of existing software platforms of genomics and transcriptomics at Bielefeld University. We demonstrate the potential of MeltDB by means of a sample experiment where we dissect the influence of three different carbon sources on the gram-negative bacterium Xanthomonas campestris pv. campestris on the level of measured metabolites. Experimental data are stored, analyzed and annotated within MeltDB and accessible via the public MeltDB web server.

Availability: The system is publicly available at http://meltdb.cebitec.uni-bielefeld.de.

Contact: hneuwege@cebitec.uni-bielefeld.de

∗To whom correspondence should be addressed.

1 INTRODUCTION

In order to achieve a complete understanding of the biological behavior of a complex system, it is essential to monitor the response of an organism to a conditional perturbation at the transcriptome, proteome and metabolome levels (Bino et al., 2004). An important step towards this goal is the integration of experimental data and results from all fields of functional genomics. Metabolomics can be regarded as the most recent contribution to this area and involves the analysis of the qualitative and quantitative collection of virtually all metabolites in the cell (the metabolome). It is furthermore most closely connected to the actual phenotype of an organism and can be linked to the genotype through knowledge about biochemical pathways and gene regulatory networks (Fiehn, 2002). Comprehensive analysis of metabolic responses has been made possible and useful with the advent of modern computational and analytical tools. The analytical procedure for metabolomics can be subdivided in three experimental sections. First, the biological samples have to be harvested without disturbance of the metabolome; the metabolites must be extracted and eventually derivatized. In a second step, the metabolites are separated by chromatographic techniques, such as gas chromatography (GC) or high-performance liquid chromatography (HPLC). In the third step, the identification and quantification of the metabolites is achieved. Hyphenated mass spectrometry is currently one of the most widely applied technologies in metabolomics, as it provides rapid, sensitive and selective qualitative and quantitative analyses (Dunn and Ellis, 2005).

The general preprocessing strategy that transforms raw datasets from hyphenated mass spectrometry to quantitative metabolite information usually encompasses noise and baseline reduction of the raw data followed by chromatographic peak detection.

After the detection and deconvolution of chromatographic peaks, the associated metabolites need to be identified. The mass spectra (MS) obtained through fragmentation by electron impact can be compared with annotated reference spectra from mass spectral databases [NIST, Golm Metabolite Database (GMD)]. If chemical ionization was performed, exact molecular masses of the mother
ions defined in, e.g. the Metlin (Smith et al., 2005) database can be used to find candidates for the chemical identity of the measured compound. Database lookup can provide evidence for the chemical identity of the metabolite but additional information, such as the retention time or retention index of the peak need to be taken into account to obtain reliable information.

Chromatographic alignment can be performed to compensate retention time drifts induced by, e.g. changes in the column performance during the measurements of an metabolomics experiment. Easily identifiable retention standards can be added to the samples yielding retention indices. Alternatively, alignments can be computed based on detected features (Robinson et al., 2007) or by finding the maximal covariance between the chromatograms (Jonsson et al., 2005). Successful alignment simplifies the assignment of corresponding unknown peaks or mass signals across multiple measurements. In order to compensate varying signal intensities due to changes in detector sensitivity or sample preparation an artificial internal standards such as ribitol can be used. Another important preprocessing step is the substitution of missing values, i.e. metabolites that could not be detected in a subset of the observed chromatograms. Simple methods for missing value estimation in metabolomics experiments are, e.g. the replacement by the mean or median of the metabolite concentrations over all samples or the imputation from nearest neighbors (Steinfath et al., 2008). After missing values have either been estimated or discarded multivariate statistics and explorative data visualizations, such as principal component analysis (PCA) and hierarchical clustering analysis (HCA) can be employed. Furthermore, pairwise correlations between the measured metabolite levels can be computed (Steuer et al., 2003) in order to gain insight into changes of the metabolomic network under varying experimental conditions.

Apart from the statistical analysis the visualization of metabolite quantities in a knowledge-based environment helps to interpret and understand experimental results. The MapMan software follows this paradigm by projection of quantitative data to metabolic pathway charts or textbook figures (Thimm et al., 2004).

It has been pointed out (Fiehn et al., 2008) that the detailed description of experimental conditions and all relevant factors of the processing of experimental data is important for the interpretation and exchange of metabolomics experiments.

Motivated by this need, a minimal set of reporting standards and best practice recommendations (Sansone et al., 2007) have emerged. Together with the definition of ontologies a standardized annotation of the experimental design and conditions has become feasible. This is of importance since experimental details are vital for the interpretation of the data and its reuse in related studies. Before reporting standards have been established, a first proposal for the standardization of metabolomics software tools was available. The ArMet (Jenkins et al., 2005) data model has been initially designed for plant metabolomics, and later has been extended to support, e.g. microbial metabolomics experiments. Based on this data model, Scholz and Fiehn (2007) have presented the SetupX database that allows for accurate description of the biological study design and accompanying meta-data in metabolomic databases. For the analysis of the experimental data, SetupX relies on the seamlessly integrated mass spectrometry database BinBase. Other available software packages for the analysis of metabolomics data are, e.g. AMDIS (Stein, 1999), met-Idea (Broeckling et al., 2006), mzMine (Katajamaa et al., 2006), MetAlign (Vos et al., 2007) or XCMS (Smith et al., 2006). The recently released TagFinder (Lüdemann et al., 2008) software does also address modern metabolomics and fluxomics studies and especially addresses non-biased metabolomic fingerprinting, footprinting and profiling experiments. Whereas all of these publicly available systems are able to perform preprocessing functions, such as peak detection, chromatogram alignments, compound identification and quantitation, the direct connection to statistical analysis tools and means to describe experimental designs and conditions is often not addressed. Besides, the connection to results from transcriptomics and proteomics experiments and genome annotation data is absent.

An open integrated system that allows to perform and evaluate preprocessing methods, annotate the experimental details and also offers means to integrate the generated datasets with functional genomics packages is to the authors’ knowledge currently not available. In this study, the MeltDB software platform is presented, which addresses this integrated approach.

2 IMPLEMENTATION AND METHODS

2.1 System design

The design of MeltDB was influenced by ArMet and the recommendations of the Metabolomics Standards Initiative (MSI) workgroup. Various classes of ArMet have been adopted but the MeltDB data model also supports user access control, a more flexible, ontology based metabolomic experiment annotation, and the possibility to integrate and parameterize preprocessing algorithms and methods that can be submitted to a compute cluster. The framework is realized using a three tier architecture consisting of a database layer, the business logic layer and the presentation layer (Fig. 1a). The O2DBI software (Linke,B., unpublished data) was used to design the data model and generate an XML document that formally describes all classes and hierarchies. Based on this formal definition, a documented application programming interface (API) was created in both Java and Perl that provides the core functionality of the MeltDB software framework. This core API is furthermore the basis of the business logic layer and provides an object relational mapping for all modeled classes of the MeltDB data model.

![Fig. 1. The three tier architecture of the MeltDB framework (a) and a symbolic representation of the main data objects and their interaction (b).](https://academic.oup.com/bioinformatics/article-abstract/24/23/2726/179514/1)

The complete data model description and the associated API documentation is available on the project web page.
2.2 Data model

An overview of the top-level classes and their interaction is given in Figure 1b. The classes Chromatogram, ChromatogramGroup and Experiment represent metabolomic experiment data. Subclasses of Chromatogram allow to distinguish the instrument-specific properties of chromatograms originating from, e.g. LC-, GC- or GC/MS measurements. ChromatogramGroup objects aggregate chromatograms from biological or technical replicate sets.

In order to be able to support the different objectives of metabolomic experiments, four specialized subclasses of Experiment have been defined in accordance with the suggestions given by the MSI.

- Targeted analysis: detection and precise quantification of a single or a small set of target compounds within a metabolome sample.
- Metabolite profiling: detection and approximate quantification of a large set of target metabolites within a metabolome sample.
- Metabolomics: detection, approximate quantification and tentative identification of as many of the compounds within a metabolome sample as possible.
- Metabolic fingerprinting: generation of a signature for a metabolome sample without regard for the individual compounds that it contains.

2.3 Experiment description

The MeltDB data model is able to describe the experimental design representing, e.g. growth conditions or extraction and sample preparation methods. Each chromatogram in MeltDB can be attributed with a list of these experimental factors which thereby also annotate the experimental conditions. The structure of the best practice recommendations of the MSI working group is integrated and can dynamically be extended. Once experimental factors have been defined in the MeltDB database, they can be reused for the annotation of multiple chromatograms.

2.4 Project management

MeltDB is a web-based application and runs on a dedicated server that provides access to various projects. A fine grained project and user management is included in order to make the system applicable to datasets that are analyzed across more than one institute with various research groups working in parallel. A generalized project management system (GPMS) is integrated and applied, which has successfully been used in other ‘omics’ software packages of the Bioinformatics Resource Facility (BRF) at Bielefeld University in dozens of international genome and transcriptome projects (Dondrup et al., 2003; Meyer et al., 2003). User authentication occurs at the login page of the MeltDB system and a personalized session is tracked. Access to experimental data is limited according to predefined rights and roles, i.e. a user with the role Guest has the right to review public datasets but does not have the right to add experimental data. Data security is ensured through MySQL permissions that are imposed on individual tables in the project databases. Higher level privileges, e.g. the right to run and modify preprocessing pipelines on raw datasets are furthermore enforced by the business logic layer of MeltDB. The careful definition of user rights and roles allows the use of the MeltDB system in a productive environment. Experimental data are by default accessible only to the owner who can then grant access to datasets and experiments to individual collaborators or user groups.

2.5 Supported input formats

MeltDB supports the established open data formats netCDF, mzData and mzXML (Orchard et al., 2007; Pedrioli et al., 2004; Unidata, 2008) that are widely used in both metabolomics and proteomics. A variety of conversion tools for these standards are available for vendor-specific file formats.1

In order to limit the data that is stored in the object relational database, unprocessed raw data in the supported data formats is stored as files and referenced together with meta information (format, acquisition date, etc.). To access MS, total ion chromatograms (TIC) or extracted ion chromatograms (EIC) of a specific chromatogram, factory classes have been implemented that generate the respective objects for the supported raw formats.

2.6 Preprocessing strategy

After the import, organization and annotation of raw chromatographic datasets in the MeltDB database, additional preprocessing steps are necessary. A possible workflow from raw data to the statistical analysis of numerical data matrices representing metabolite concentrations is depicted in the Supplementary Material. This figure does also present the functionality provided by or integrated into MeltDB. In order to be able to support a wide variety of analytical platforms, no strict preprocessing pipeline is enforced by MeltDB. The system does in fact allow to complement results from vendor-specific software systems such as ChromaTOF (LECO) or Xcalibur (Thermo Finnigam) with available academic tools that, e.g. perform peak detection (Meta-B) or chromatogram alignments (XCMS). The most simple approach to use MeltDB would be to import peaks that have been detected, quantified and identified by, e.g. ChromaTOF and use them as basis for further statistical analysis and data integration. Nonetheless, integrated methods can be employed by the user to, e.g. detect peaks missed by vendor-specific software systems or to compute chromatographic alignments. MeltDB does by design allow to combine, compare and evaluate the results of different preprocessing methods.

2.7 Tool concept

MeltDB features a flexible tool concept that facilitates the integration of existing open software packages from computational metabolomics such as, e.g. XCMS (Smith et al., 2006), or Meta-B.2

It might become necessary to adopt the parameterization of these tools to the datasets under study. Thus, the generic Tool concept allows to store multiple parameterizations of single methods and applications in the MeltDB database. The experimenter can select a suitable tool instance for the preprocessing of his chromatograms or experiments and submit the actual computation to a compute cluster. MeltDB utilizes the Distributed Resource Management Application API (DRMAA)3 as a high-level API specification for the submission and control of jobs to Distributed Resource Management Systems (DRMS). In order to connect the computed results with peaks, chromatograms or experiments stored in the MeltDB database, Observation and Annotation objects are created.

An observation that is associated to a chromatographic peak represents, e.g. the computed retention index or a match to a MS from the GMD database. Whereas observations may contain hypothetical and contradicting information, annotations are used to aggregate approved information. Observations and annotations are linked to the creating tool, thus it is ensured that the generated results are reproducible and transparent. The general concept of tools, observations and annotations is presented in Figure 1b. All tools have strictly typed input and output objects and can therefore be combined to pipelines in MeltDB that cover the necessary preprocessing steps.

2.7.1 Implemented tools

Whereas the focus of this article is on the description of the MeltDB system design, two processing methods for raw data that are realized using functionality of the MeltDB API are presented exemplarily in this section.

- Metabolite identification by mass spectral database lookup.
- Retention Index computation and filtering.

1sashimi.sourceforge.org


The following function for \( D \) between the vector representation of database and query spectra is used as a similarity measure and is calculated as follows:

\[
D = \sum_{j} \hat{I}_{\text{library},j} \hat{I}_{\text{query},j}
\]

(1)

where \( \hat{I}_{\text{library},j} \) is the normalized intensity of the \( j \)-th \( m/z \) bin of the library spectrum, and \( \hat{I}_{\text{query},j} \) that of the matching bin of the query spectrum. The length of both vectors is normalized to 1, thus the value of \( D \) is always between 0 and 1, with the latter value indicating identical spectra (Lam et al., 2007).

Mass spectral databases in Structure Data Format (SDF), such as the GMD (Kopka et al., 2005) can be parsed with MeltDB functionality and are used to generate observations for the chemical identity of chromatographic peaks. The mass spectrum located at the peak apex is therefore extracted from the raw data and compared against all database spectra. The similarity \( D \) is used to select database spectra above a given threshold \( 0 < t < 1 \) for which the associated information, such as compound name, CAS number or KEGG Compound ID are again connected to the peak via observations. The complete functionality is encapsulated in the Tool::PeakIdentification::Cosine class of the MeltDB API.

The GMD contains MS for known compounds associated with information on retention time and Kovat’s Retention Indices (Ettre, 1993). By combining the peak identification tool with a specialized retention index library, MeltDB is able to identify peaks representing retention indices (e.g., Dodecane or Pentadecane) that have been added to a biological sample. Relative to the retention time of these identified standards, retention indices can be computed for all other peaks detected in a chromatogram (RI\(_{\text{peak}}\)). By combining the deviation of retention indices from the database RI\(_{\text{db}}\) with the mass spectral similarity, the number of potential database matches is reduced. The following function for \( D' \) is used in MeltDB as an additional filter to exclude matches if the difference of RI\(_{\text{peak}}\) and RI\(_{\text{db}}\) becomes too large.

\[
D' = \frac{D}{\epsilon^{\frac{RI_{\text{peak}} - RI_{\text{db}}}{2}}}
\]

(2)

The parameter \( \epsilon \) is used to adjust the tolerated deviations of the retention indices according to the attributes of the chromatographic instrument which can be defined for tool instances of the class Tool::PeakIdentification::RI in the MeltDB API.

### 2.7.2 Importer

For preprocessing methods that cannot be employed in an open source environment, integration has been achieved through specialized importing functionality using the previously described tool concept. For the following file formats, importers have been implemented:

- AMDIS reports
- Thermo Xcalibur reports (XLS format)
- LECO ChromaTOF reports
- MassHunter reports (Text format)
- ChemStation peak lists (Text format)

The importers transform information on the chromatographic peaks predicted by external software tools into the MeltDB data model. Information associated to peaks such as chemical identity, peak area and intensity are represented using observations and annotations. The extensible importer concept realized in the MeltDB data model makes it easy to import additional text-based vendor-specific formats. Together with the support for netCDF, mzXML and mzData metabolomics experiments conducted on various GC and LC/MS instruments can be analyzed using MeltDB.

### 2.8 User interface

The functionality of MeltDB can be accessed through a platform-independent web interface. A Perl CGI script running on an Apache web server dynamically creates the HTML content and manages the authentication and the ongoing user sessions. The interactivity of the web application is increased through the use of mod_perl, JavaScript and AJAX which also results in fast access to all objects stored in the MeltDB database. In order to obtain a flexible and extendable web interface, the Model-View-Controller design pattern (Gamma et al., 1995) is employed for the generation of the actual HTML content. Generic views applicable to all objects in the MeltDB database provide a tabular representation of, e.g., the annotations associated to a peak. Specialized views realizing the same interface have been added to the system. Thus, the TICs and MS of MeltDB objects, such as chromatograms, peaks, or complete experiments can be represented using graphs and diagrams. Some of these views and visualization methods provided by the web interface are presented in Figures 2 and 3.

Apart from the visualization of raw data and results, investigators can also employ the preprocessing and import functionality included in MeltDB via the web interface. The annotation of ‘unknown’ peaks with respect to chemical identity can be performed by the researcher via the web interface. If no matching MS can be found in the GMD or NIST databases, the integrated methods for the computation of mass spectral similarity can also be applied to find matching peaks in other publicly available chromatograms organized in MeltDB. If an annotated peak with matching MS, retention time or index could be detected, the information may be inherited in order to annotate the yet unknown peak. Thereby the number of ‘unknowns’ can be reduced step by step.

### 2.9 Visualization methods

The availability of standardized interfaces to the different raw file formats provided by the MeltDB API allows to realize generic visualizations, such as TIC views and chromatogram alignments for whole experiments as shown in Figure 2. All visualizations are enriched with information on detected peaks in the MeltDB web interface. Interactive access to the underlying information stored in associated observations and annotations is provided via the web interface.
Access to R functionality is realized in MeltDB using the RSPerl modules. The data matrix resulting from a normalization of peak areas or intensities of linkage method are applied. Chromatograms (also allows to visualize the results of HCA on both metabolites and metabolites can, e.g. be visualized using boxplots (a). The web interface also allows to visualize the results of HCA on both metabolites and chromograms (b). The Euclidean distance function and the complete linkage method are applied.

2.10 Statistical analysis
Freely available R packages from BioConductor offer a multitude of analysis functions and visualizations developed for functional genomics datasets. Access to R functionality is realized in MeltDB using the RSPerl modules.4 MeltDB database objects are dynamically converted to data objects in R in a standardized manner which results in an integrated and efficient way to statistically analyze metabolomic experiments stored in MeltDB. The cumbersome and error prone conversion of data tables from proprietary software packages into a format that a statistic software framework can interpret and analyze is avoided. Experimental factors assigned to a MeltDB experiment are furthermore projected to the R representations of the data and allow to interpret the visualizations together with the experimental design. The data matrix resulting from a normalization of peak areas or intensities of identified compounds using an internal standard (e.g. ribitol) and dry-weight information is used for the box plot visualization presented in Figure 3a. All statistical methods can be executed using the web interface and additional criteria, such as the treatment of missing values, the scaling of the values and the exclusion of certain metabolites or whole chromatograms can be controlled by the user. The visualization of independent component analysis (ICA; Hyvärinen and Oja, 2000) and PCA presented in the Supplementary Material, the results of a HCA (Fig. 3b) or the results of a pairwise correlation analysis of the pool sizes of the measured metabolites can easily be generated and exported in either PDF or PNG format. The following list shows the statistical and explorative analysis methods currently available in MeltDB:

- Student’s t-test
- Analysis of Variances (ANOVA)
- HCA
- PCA
- ICA
- Metabolite Correlation Analysis
- Volcano Plots

Additional analysis and visualization functionality provided through BioConductor packages or R can easily be added to MeltDB. Functionality can either be realized using the MeltDB Tool concept or data visualizations generated by R can be embedded in the MeltDB web interface.

2.11 Data integration
Data integration in MeltDB is achieved on various levels. The KEGG compound database (Kanehisa et al., 2006) is regularly imported through direct access to the KEGG FTP server. Relevant terms and relations of the compounds are directly represented in the MeltDB database model. The current version of the KEGG compound database contains ≈15,000 entries. These entries act as a controlled vocabulary in MeltDB for compounds that are relevant for biological systems. References to other metabolite databases (CheBI, CAS) and the connection to metabolic pathways are included in the MeltDB database representation.

The connection to existing functional genomics systems, such as Emma2 (Dondrup et al., 2003) for transcriptomics experiments or GenDB (Meyer et al., 2003) for genome annotation data is realized via SOAP-based web services. As demonstrated recently (Neuweger et al., 2007), web services are an efficient means for the integration of heterogeneous frameworks.

Since we use the KEGG compound database as a controlled vocabulary for compounds and thereby obtain a connection of metabolites to pathways, enzymes and genes, it is straightforward to link metabolic experiments with existing genome projects stored in GenDB. Gene annotations containing EC numbers are requested via the GenDB web service and are used to enrich metabolic pathway representations from the KEGG database together with qualitative information on detected metabolites (Fig. 4). As soon as combined experiments on both metabolites and transcript measurements are available in the Emma2 system, the presented web service approach can be used to connect the respective datasets.

3 APPLICATION EXAMPLE AND RESULTS
We contribute a functional web-based system for the storage, administration, analysis and integration of metabolomics datasets. The advanced visualization and analysis features of MeltDB are presented in more detail by the following metabolomic experiment conducted on the bacterium Xanthomonas campestris pv. campestris B100.

3.1 Xcc B100 grown on three different carbon sources
The genus Xanthomonas mainly consists of phyto-pathogens of wide-host range. On their ability to infect different host-plants the genus is subdivided in species and pathovars, a classification system that has also been verified by 16S rDNA sequence analysis. One of these species is X. campestris pv. campestris (Xcc), the causal agent of black rot disease in crucifers. Xcc secretes numerous lytic enzymes and genes, it is straightforward to link metabolic experiments with existing genome projects stored in GenDB. Gene annotations containing EC numbers are requested via the GenDB web service and are used to enrich metabolic pathway representations from the KEGG database together with qualitative information on detected metabolites (Fig. 4). As soon as combined experiments on both metabolites and transcript measurements are available in the Emma2 system, the presented web service approach can be used to connect the respective datasets.

3 APPLICATION EXAMPLE AND RESULTS
We contribute a functional web-based system for the storage, administration, analysis and integration of metabolomics datasets. The advanced visualization and analysis features of MeltDB are presented in more detail by the following metabolomic experiment conducted on the bacterium Xanthomonas campestris pv. campestris B100.

3.1 Xcc B100 grown on three different carbon sources
The genus Xanthomonas mainly consists of phyto-pathogens of wide-host range. On their ability to infect different host-plants the genus is subdivided in species and pathovars, a classification system that has also been verified by 16S rDNA sequence analysis. One of these species is X. campestris pv. campestris (Xcc), the causal agent of black rot disease in crucifers. Xcc secretes numerous lytic enzymes and genes, it is straightforward to link metabolic experiments with existing genome projects stored in GenDB. Gene annotations containing EC numbers are requested via the GenDB web service and are used to enrich metabolic pathway representations from the KEGG database together with qualitative information on detected metabolites (Fig. 4). As soon as combined experiments on both metabolites and transcript measurements are available in the Emma2 system, the presented web service approach can be used to connect the respective datasets.

3 APPLICATION EXAMPLE AND RESULTS
We contribute a functional web-based system for the storage, administration, analysis and integration of metabolomics datasets. The advanced visualization and analysis features of MeltDB are presented in more detail by the following metabolomic experiment conducted on the bacterium Xanthomonas campestris pv. campestris B100.

3.1 Xcc B100 grown on three different carbon sources
The genus Xanthomonas mainly consists of phyto-pathogens of wide-host range. On their ability to infect different host-plants the genus is subdivided in species and pathovars, a classification system that has also been verified by 16S rDNA sequence analysis. One of these species is X. campestris pv. campestris (Xcc), the causal agent of black rot disease in crucifers. Xcc secretes numerous lytic enzymes and genes, it is straightforward to link metabolic experiments with existing genome projects stored in GenDB. Gene annotations containing EC numbers are requested via the GenDB web service and are used to enrich metabolic pathway representations from the KEGG database together with qualitative information on detected metabolites (Fig. 4). As soon as combined experiments on both metabolites and transcript measurements are available in the Emma2 system, the presented web service approach can be used to connect the respective datasets.

3 APPLICATION EXAMPLE AND RESULTS
We contribute a functional web-based system for the storage, administration, analysis and integration of metabolomics datasets. The advanced visualization and analysis features of MeltDB are presented in more detail by the following metabolomic experiment conducted on the bacterium Xanthomonas campestris pv. campestris B100.

3.1 Xcc B100 grown on three different carbon sources
The genus Xanthomonas mainly consists of phyto-pathogens of wide-host range. On their ability to infect different host-plants the genus is subdivided in species and pathovars, a classification system that has also been verified by 16S rDNA sequence analysis. One of these species is X. campestris pv. campestris (Xcc), the causal agent of black rot disease in crucifers. Xcc secretes numerous lytic enzymes and genes, it is straightforward to link metabolic experiments with existing genome projects stored in GenDB. Gene annotations containing EC numbers are requested via the GenDB web service and are used to enrich metabolic pathway representations from the KEGG database together with qualitative information on detected metabolites (Fig. 4). As soon as combined experiments on both metabolites and transcript measurements are available in the Emma2 system, the presented web service approach can be used to connect the respective datasets.

4 www.omegahat.org
enriches and an exopolysaccharide (EPS) to facilitate its pathogenic and saprophytic life style. Not only the pathogenic properties put this rod-shaped, gram-negative bacterium in the focus of interest, but also the exopolysaccharide, so-called xanthan gum, produced by Xcc is of economical importance. Xanthan gum finds a variety of industrial uses as stabilizer in foods, cosmetics, paints and as bio-lubricant in oil drilling. As biotechnological product it therefore takes 1% of the world market shares and is estimated to have an annual turnover of US $400 million (2004). The strain used in this work is X. campestris pv. campesstris B100 (Xcc B100). The proteome of this strain has been widely researched and only recently the genome of Xcc B100 has been published (Vorhölter et al., 2008). But the EPS production and regulation still needs to be thoroughly scrutinized on the metabolomic plain. Under certain conditions Xcc is able to channel up to 80% of the available carbon source into the xanthan gum production.

Xanthomonas campestris pv. campesstris B100 was cultivated in Vincent-Minimal Medium (VMM) in shaken flasks. The medium was supplemented with 1% (w/v) of glucose, mannitol and succinate, respectively, and two biological replicates each. Samples were taken in the early stationary phase by rapid centrifugation. After freeze-drying the biomass was disrupted with a bead-beater and the hydrophilic metabolites extracted with 80% methanol containing 10 μM ribitol (internal standard). Derivatization and GC-MS measurements were conducted as described previously (Barsch et al., 2004). The resulting chromatograms were converted to netCDF format using functionality of Xcalibur and imported into the MeltDB system. Annotation of the chromatograms was done in accordance with the recommendations of the MSI and the chromatograms were organized in replicate groups according to their carbon source in MeltDB. Peak detection, identification and quantification was performed using the Xcalibur software. The Xcalibur importer tool of MeltDB was used to transfer these results into the MeltDB data model and link identified metabolites to the KEGG compound database.

An initial comparative inspection of the raw datasets was performed using the Experiment TIC plot provided by the MeltDB web interface (Fig. 2b). Although retention time deviations of several seconds were observed between the chromatograms, most of the predicted peaks of the Xcalibur preprocessing method are present in all samples.

After the MeltDB chromatogram alignment was performed and visualized, deviations in peak intensities between the different replicate groups became evident (Fig. 2c). To analyze the differences in more detail, peak areas were normalized in each chromatogram relative to the area of the ribitol peak (set to 100). The data matrix computed within MeltDB was used for further statistical analysis such as, e.g. simple boxplot visualizations shown in Figure 3a.

The heatmap visualization of a hierarchical clustering on both metabolites and chromatograms shown in Figure 3b highlights that the replicates of the three groups (glucose, mannitol and succinate) cluster together. The dendrogram associated to the chromatograms included in this feature also shows that the glucose and mannitol groups are more similar to each other than to the succinate group. The integrated PCA of MeltDB allows a closer look at the metabolites responsible for the clustering properties of the samples. As can be seen in the PCA and ICA visualizations (Supplementary Material), the replicate groups again cluster together. Higher abundances of 3-phospho-glycerate, 1,2-aminoadipate and glycerone phosphate in the glucose and mannitol approach, and the higher amounts of 1-leucine and uracil in the succinate approach seem to be the main reason why the mannitol and glucose group differ from the succinate group. The main differences between the glucose and mannitol group are found in the higher metabolic pools of gluconic acid and d-erythrose 4-phosphate, when glucose is the sole carbon source. Boxplots of the normalized intensities for these metabolites are depicted in Figure 3a.

From the heatmap and PCA visualizations one could already assume that there are major differences between the succinate group and the glucose and respectively mannitol group, whereas there still seem to be some similarities in the way mannitol and glucose are metabolized.

The PCA indicates a high abundance of gluconic acid and d-erythrose 4-phosphate in the glucose group. Since both are direct or indirect intermediates of the pentose phosphate pathway (Fig. 4), one may assume that glucose is not only catabolized via the glycolysis and pentose and glucuronate interconversions, but in contrast to the mannitol group via this pathway also. On the other hand there were no differences between the glucose and mannitol group in the amount of 3-phospho-glycerate and 2-phospho-glycerate, which indicates that Xcc B100 metabolizes mannitol and glucose with seemingly the same mass flux via the glycolysis. The higher abundance of glycerone phosphate in the mannitol group again shows that mannitol degradation not only differs from glucose metabolism in the low utilization of the pentose phosphate pathway, but also in the higher usage of the glucuronate interconversions, since glycerone phosphate is an intermediate found in the periphery of this pathway. Since succinate is directly internalized into the citrate cycle, it is not catabolized via any of the above pathways. Rather glucose is anabolized via gluconeogenesis, which explains the higher number of differences of the succinate group to the other two.
These findings are facilitated by MeltDB through a mapping of the detected metabolites onto the metabolic pathway maps provided by the integrated KEGG database. As MeltDB is also connected with the functional genomics packages GenDB and Emma2 via web services and the BRIDGE layer (Goesmann et al., 2005), current gene annotation data from the XCC annotation project is dynamically added to the pentose phosphate pathway (Fig. 4). Thus, MeltDB supports the analysis of metabolic experiments in a knowledge-based environment and simplifies the interpretation of experimental results in a biological context.

4 CONCLUSION

MeltDB was developed as a platform-independent software for the analysis and integration of metabolomics experiments. The systems functionality can be easily accessed through a user-friendly web interface. As presented in the previous section, MeltDB can be applied for the description and analysis of metabolomic experiments. Researchers can furthermore benefit from the integration of the MeltDB system with other functional genomics packages and databases. A public project contains the experiment presented in the previous chapter and can be accessed via the web interface using the guest login.

The existing MeltDB projects currently host more than 30 experiments predominantly from GC/MS measurements. Nonetheless, the system is not limited to this technology by design. Through the integration of, e.g. the XCMS package preprocessing functionality for LC/MS experiments is available.

MeltDB controls project- and user-specific access to metabolomics experiments. With the established infrastructure at hand, the analysis of metabolomics experiments is being simplified and standardized for the researcher. At the same time, the generic nature of the system and standardized access to various raw data formats through the MeltDB API allows to include and evaluate novel methods and algorithms for the preprocessing of metabolomic datasets. These can easily be evaluated in the context of existing results and expert annotations. The connection to a compute cluster by DRMAA interface provides the scalability necessary for high-throughput analysis.

ACKNOWLEDGEMENTS

H.N. and T.W. would like to thank the International Graduate School in Bioinformatics and Genome Research for providing financial support. M.D. acknowledges financial support by the BMBF (grant 0313805A ‘GenoMik-Plus’). The authors wish to thank the BRF team for expert technical support.

Funding: BMBF, Germany, in the frame of the QuantPro initiative (grant 0313812).

Conflict of Interest: none declared.

REFERENCES


