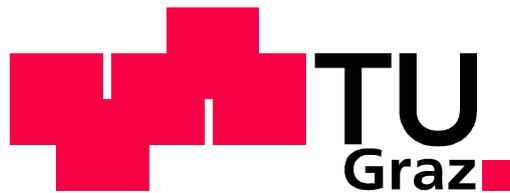


BIOINFORMATICS PLATFORM FOR LARGE
SCALE PROTEOMICS LIQUID
CHROMATOGRAPHY TANDEM MASS
SPECTROMETRY DATA

JÜRGEN HARTLER



Doctoral Thesis

Graz University of Technology
Institute for Genomics and Bioinformatics
Petersgasse 14, 8010 Graz, Austria

Graz, April 2007

Abstract

Mass spectrometry has emerged to become a state-of-the-art methodology for the analysis of the proteome. In order to manage and extract valuable information from mass spectrometry data, a computational data management platform is indispensable. Furthermore the standardized presentation of data is of the utmost importance. At present however, no proteomics bioinformatics platform exists that fulfills the Minimum Information About a Proteomics Experiment (MIAPE) standard and that provides all the features needed for the effective analysis of mass spectrometry experiments.

In view of this we have developed a versatile MIAPE compliant platform for the management and analysis of proteomics mass spectrometry experiments: the MAss SPECTRometry Analysis System (MASPECTRAS). This platform is web-based with a database back-end and relies on the Java 2 Enterprise Edition development platform. The platform is scalable and enables the outsourcing of computationally intensive tasks to a computing cluster. The data model captures data concerning experimental design and at all other subsequent steps leading up to evaluation and result export. The analysis process relies on the output of the major mass spectrometry search engines SEQUEST, Mascot, SpectrumMill, X!Tandem and OMSSA. The data is run through an automated analysis pipeline in the following four steps: 1) import and parsing to a MIAPE compliant representation; 2) validation; 3) protein clustering; 4) peptide quantification. For manual analysis MASPECTRAS provides unique features that include the merging of results originating from different search engines, a chromatogram viewer and export to PRIDE-XML format which can be uploaded directly in a public repository. The functionality of the whole system is furthermore embedded in a multi-user environment providing controlled user access. The platform has been evaluated using large-scale and quantitative data.

The MASPECTRAS platform offers researchers an environment for the rapid analysis of large-scale proteomics experiments. Due to its modular design it is flexible enough to easily accommodate future changes in proteomics data management.

Keywords: proteomics, management platform, tandem mass spectrometry (MS/MS), MIAPE, large-scale analysis, PRIDE, AndroMDA, J2EE

Publications

This thesis was based upon the following publications, as well as upon unpublished work:

Papers

Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, Körner E, Rader R, Schmidt A, Mechtler K, and Trajanoski Z: **MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data.** *submitted to BMC Bioinformatics*

Maurer M, Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, Prokesch A, Scheideler M, and Trajanoski Z: **MARS: microarray analysis, retrieval, and storage system.** *BMC Bioinformatics* 2005, **6**:101.

Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, and Trajanoski Z: **PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.** *Nucleic Acids Res* 2005, **33**:W633-W637.

Hackl H, Maurer M, Mlecnik B, Hartler J, Stocker G, Miranda-Saavedra D, and Trajanoski Z: **GOLDDb: Genomics of lipid-associated disorders database.** *BMC Genomics* 2004, **5**:93.

Conference Proceedings

Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, Körner E, Mechtler K, and Trajanoski Z: **Management and Analysis of Proteomics LC-MS/MS Data.** Invited talk at: *Fourth International Symposium of the Austrian Proteomics Platform. Seefeld in Tirol.* 28.01.2007

Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, Körner E, Fuchs T, Mechtler K, and Trajanoski Z: **MASPECTRAS: Web-based System for Storage, Retrieval, Quantification and Analysis of Proteomic LC MS/MS Data.** – Poster award at: *Third International Symposium of the Austrian Proteomics Platform, Seefeld in Tirol.* 16.01.2006

Hartler J, Thallinger GG, Sturm A, Burkard TR, Körner E, Fuchs T, Mechtler K, and Trajanoski Z:
MASPECTRAS: Web-basiertes Datenbanksystem zur Verwaltung von Proteomik-Daten. *Lange Nacht der Wissenschaft, Vienna.* 01.10.2005

Hartler J, Thallinger GG, Sturm A, Burkard TR, Körner E, Fuchs T, Mechtler K, and Trajanoski Z:
MASPECTRAS: Web-based System for Storage, Retrieval, and Analysis of Proteomic LC MS/MS Data. *HUPO 4th Annual World Congress, Munich.* 29.08.2005

Hartler J, Thallinger GG, Morandell S, Huber LA, Mechtler K, and Trajanoski, Z.: **Development of a web based mass spectrometry analysis system.** *ÖGBMT Symposium. Graz.* 12.11.2004

Contents

LIST OF FIGURES	V
LIST OF TABLES	VI
1. INTRODUCTION	1
1.1 Background.....	1
1.2 Objectives.....	4
2. METHODS	5
2.1 Relative protein quantification.....	5
2.2 Standards in proteomics	6
2.3 Software Technology	8
2.4 Experimental procedures.....	21
3. RESULTS.....	22
3.1 Extensions to the AndroMDA environment.....	22
3.2 MAAss SPECTRometry Analysis System (MASPECTRAS).....	29
3.2.1 MIAPE compliance.....	29
3.2.2 Analysis pipeline.....	31
3.2.3 Data import	33
3.2.4 Data validation	36
3.2.5 Protein clustering	37
3.2.6 Peptide quantification	37
3.2.7 Visualization – views on the data.....	39
3.2.8 Visualization – additional visualization tools	42
3.2.9 Exporting.....	47
3.3 Validation study using large-scale data.....	48

3.4 Validation study using quantitative data	48
4. DISCUSSION	50
BIBLIOGRAPHY	55
APPENDIX A – MASS SPECTROMETRY	64
APPENDIX B – DATABASE SEARCH ENGINES	68
APPENDIX C – MASPECTRAS SCHEMES	70
GLOSSARY	74
ACKNOWLEDGEMENTS	77
PUBLICATIONS	78

List of Figures

1	Schematic overview of the relative quantification process	6
2	Schema of a multi-tiered application	9
3	Class diagram of the session façade design pattern	12
4	JSP model 1 architecture	14
5	JSP model 2 architecture	15
6	Workflow of model driven architecture	16
7	Overview of the AndroMDA build process used for the development of a J2EE application	18
8	UML-model of a report bean for AndroMDA	24
9	UML diagram for a fully functional report page	25
10	Two web-pages generated by AndroMDA	26
11	Sharing in the AndroMDA environment	28
12	MIAPE compliant input mask for 2 dimensional gels and navigation tree	30
13	MIAPE compliant input masks for liquid column chromatography and for a band of a 1-dimensional gel electrophoresis experiments	31
14	MASPECTRAS analysis pipeline	32
15	Schema for the storage of PTMs	34
16	Generic protein sequence database administration	35
17	Schema for the storage of proteins	36
18	Peptide view	40
19	Protein view and clustered protein view	41
20	Merging of results in MASPECTRAS	42
21	Jalview alignment editor	42
22	Spectrum viewer of MASPECTRAS	44
23	Chromatogram viewer of MASPECTRAS	45
24	Quantitative comparison module	46
25	Excerpt of a PRIDE-XML document generated by MASPECTRAS	47

List of Tables

1	Comparison of the results of a large-scale study performed with MASPECTRAS with the published results	48
2	Quantitative evaluation of ICPL-labeled probes by MASPECTRAS, MSQuant and PepQuan	49
3	Comparison of MASPECTRAS to other proteomics tools	51

Chapter 1

Introduction

1.1 Background

Genomics is a relatively young discipline that approaches genetics at the molecular level and on a genome-wide scale. Despite this youth however, it had nevertheless developed into a major area of biological research by the end of the 20th century largely due to a perceived potential to provide a better understanding of complex diseases such as cancer and diabetes mellitus. Following the accumulation of vast amounts of DNA sequences in databases and the completion of the human genome sequencing project, researchers however realized that complete genome sequences in databases are alone not sufficient to elucidate biological function. “There is no strict linear correlation in respect of the genes of a genome and the corresponding proteins or ‘proteome’ of a cell” [1]. Proteomics in contrast has great potential to elucidate function since it studies gene products which are the active agents in the cell. Indeed proteomics is directly relevant to the process of drug design “as almost all drugs are directed to proteins” [1].

One of the most important applications of proteomics is the characterization of posttranslational protein modifications (PTMs). “Proteins are known to be modified posttranslationally in response to a variety of intracellular and extracellular signals” [2]. Protein phosphorylation for example plays a major role in the regulation of many cellular processes in eukaryotes such as signaling, the deregulation of which can result in oncogenesis [3]. Since only a certain percentage of the proteins involved in a biological signaling pathway are modified an ability to selectively quantify modified and unmodified proteins involved in signaling is important for the better understanding of biological processes. This process, known as protein expression proteomics, deals with “the quantitative study of protein expressions between samples that differ by some variable” [2] and focuses on the comparison of the entire proteome or the subproteome between samples.

Advancements in proteomics have been driven by the progress made in protein technologies. In parallel with the development of improved techniques for the separation and isolation of proteins, mass spectrometry (MS) has emerged as the state-of-the-art technology for the acquisition of information about proteins. Mass spectrometry has essentially replaced classical technique of Edman

sequencing, developed in 1949, because it is much more sensitive, can process protein mixtures and offers much higher throughput [1]. Mass spectrometry reveals structural information about proteins, through peptide masses or amino acid sequences [2]. “No method or instrument exists that is capable of identifying and quantifying the components of complex protein samples in a simple, single-step operation. Rather, individual components for separating, identifying and quantifying the polypeptides as well as tools for the integrating and analyzing all the data must be used in concert” [4]. MS-based proteomics has established itself as an indispensable technology to interpret information encoded in genomes, to reveal sequence- and posttranslational information, and to quantify the components.

One of the most common techniques is liquid chromatography tandem mass spectrometry (LC-MS/MS) which “is that method that is today at the core of MS-based proteomics” [4]. Although used as early as 1992 [5], widespread usage of this method followed the introduction of the automated large-scale approach of peptide fragmentation fingerprinting (PFF) using search engines [6]. PFF analyzes peptides obtained from digested mixtures of intact proteins [7]. Generally, the peptides are separated by a reverse phase liquid chromatography (LC) column to reduce sample complexity before loading into the mass spectrometer. In a first run (MS) the peptides are separated according to their mass to charge ratio (m/z). Single ions (normally the ones with the highest intensity) are selected from the mixture and fragmented. In a second run (MS/MS) the fragments are separated according to their mass to charge ratio. The derived pattern is an indicator of the amino acid sequence (for a more detailed description see Appendix A). For the identification of the corresponding amino acid sequence, there are several PFF search engines available. These are based on the same principle but use different scoring algorithms. Biological protein databases are enzymatically digested *in silico*. Candidate sequences are selected for each spectrum on the basis of the molecular weight of the peptide. The *in silico* calculated fragmentation pattern is then compared to that of the spectrum and the accordance is expressed as a score [8] (for a more detailed description see Appendix B).

The first available PFF search algorithm was SEQUEST [9,10], followed by Mascot [11]. In recent years, most effort has been invested, in the improvement of mass spectrometers on the one hand (improved sensitivity, resolution, mass accuracy, information rich spectra) and in the development and improvement of search algorithms (especially their reliability) on the other [12-24]. In addition, studies of the advantages and disadvantages of the scores have been performed [6,25]. These ranked the algorithms according to their sensitivity and specificity but did not suggest the best algorithm to use. They rather proposed the idea of consensus scoring [25] using a combination of different search algorithms for better results (e.g. the combination of an algorithm with high sensitivity and one with high specificity).

Nevertheless, proteomics research is not restricted to the identification of peptides and the corresponding proteins. The data must be further processed (e.g. clustering of similar proteins, conversion to other file formats) and the quantitative information of the experiments extracted. To address this issue several bioinformatics tools have been developed [26-29]. The large number of

applications in this area however leads to confusion and incompatibilities and demands significant bioinformatic understanding from proteomics researchers. There is a clear need to improve the consistency and integrity.

Another important issue is the management and standardized distribution of proteomics data, where a uniform reusable presentation is desirable. The microarray community jointly defined the critical information necessary to effectively describe a microarray experiment and developed the Minimum Information About a Microarray Experiment (MIAME) standard [30]. Subsequently, MIAME was adopted by scientific journals and several software platforms supporting MIAME have been developed [31,32]. The principles underlying MIAME have resonated beyond the microarray community. The HUman Proteome Organization (HUPO) [33] established the Proteomics Standards Initiative (PSI) [34,35] in order to provide a similar level of standardization for proteomics. PSI defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification [36]. With reference to MIAME they defined the Minimum Information About a Proteomics Experiment (MIAPE) standard for proteomics [36-43]. The overall MIAPE standard is composed of several parts, subject to ongoing development, that describe steps for the sample processing before entering the mass spectrometer (gels, chromatography, etc), information about the specific mass spectrometer used and the settings and results for the database searches. Some of these consist only of working drafts which can be rapidly changed. As well as the MIAPE standard, large repositories for proteomics data have emerged, i.e. the Proteome Experimental Data Repository (PEDRo) [44] or the PRoteomics IDentifications database (PRIDE) [45].

Many organizations have already realized that the analytical power of classical search engines such as SEQUEST and Mascot is insufficient for large scale studies. This has stimulated the development of additional commercial and freely available applications [46-60]. To address these problems some of these applications have placed more emphasis on data analysis whilst others focused on the distribution of data and conformity to standards. The uniform proteomics MS/MS [46] analysis platform for example consists of a package of analysis tools which interact with each other using open XML file formats for data at the raw, peptide and protein level. In contrast, PROTEIOS [52] focuses on a standard-conform presentation of the data and on exporting features. A further limitation is that most of the available tools can only be applied to a single LC-MS/MS search run, totally contradicting the idea of consensus scoring. Only by comparing separate experiments (e.g. cells at different states, tumor cells versus normal cells) can precious information concerning complex diseases be unraveled. Stated again, each of these applications addresses only parts of the requirements for a uniform all-in-one solution for the analysis and management of proteomics data. A proteomics platform, enabling the analysis and the standardized storage of data for distribution thus far missing, would be of great value for proteomics research.

1.2 Objectives

The main goal of this thesis was to develop a uniform, scalable and extendable proteomics platform for the management and analysis of LC-MS/MS data. The resulting system should cope with heterogeneous data types and file formats and contain a set of well-established analysis and evaluation tools integrated in one application. It should be MIAPE compliant while remaining sufficiently flexible and component-based to cope with future changes and adaptations in this rapidly evolving research field. Furthermore it should provide well-defined user and data interfaces and fine-grained access to the data. The requirements for such a system comprise of the following points:

- Import compatibility
 - a) Results of the most commonly used search engines SEQUEST, Mascot, SpectrumMill [61], X!Tandem [20], and OMSSA [23]
 - b) Raw data in standardized format (mzXML, mzData)
 - c) Biological protein sequence databases
- Additional tasks performed by the platform:
 - a) Validation features
 - b) Clustering of similar proteins
 - c) Quantification of the results
- Presentation of the data:
 - a) Summary view on the identified proteins and peptides
 - b) Merged view of results of different search engines
 - c) Querying and filtering of the data according to the users' needs
 - d) Export of the results to an external file formats for dissemination
 - e) Graphical tools for manual validation
- Performance issues:
 - a) Support for large datasets (raw files around 200MB and search results up to 700MB)
 - b) Reasonable response times
 - c) Asynchronous processing of long lasting tasks
- Data storage:
 - a) Guarantee for integrity
 - b) Efficient access
 - c) Fine-grained data access
 - d) Independence of database management system (RDBMS) specific features
 - e) MIAPE compliance
 - f) Easy possibilities for maintenance and extensions

Chapter 2

Methods

2.1 Relative protein quantification

Mass spectrometry delivers quantitative information about the ions involved [4], but the extraction is not a single step procedure (see figure 1). Han et al. [27] and Gygi et al. [62] showed that the relative abundance of a peptide can be determined by comparing the obtained peak areas of chromatograms. The chromatogram itself is not part of the direct output of a mass spectrometer. It can be calculated by summing up the intensities of the individual MS spectra, using its inherent quantitative information. The MS/MS spectra are not of interest, because they are only results of fragmented peaks from the MS spectra (see Appendix A). The whole chromatogram is a measure of mass influx in the mass spectrometer. To determine the chromatogram of a peptide, only contributions of a specific mass can be accepted. Since the peptides are occurring at different charge states and the mass spectrometer measures the mass indirectly by the mass to charge ratio, multiple chromatograms have to be taken into account. The measured MS peaks belonging to one peptide can deviate from the ideal m/z value according to the mass spectrometers resolution and accuracy. Therefore a m/z range has to be used for the calculation of the chromatogram which is dependant on the used mass analyzer. For the integration of the peak area of the curve additional smoothing is necessary. The resulting peak area is a measure for the relative abundance of the peptide and does not correspond to the total mass of the peptide in the biological sample. This can be explained by the ionization (and the resulting signal) which is extremely dependant on the physiochemical properties of the peptide [63]. Therefore only a relative quantification of similar or the same peptides is possible without additional aid of labeling techniques [64-67].

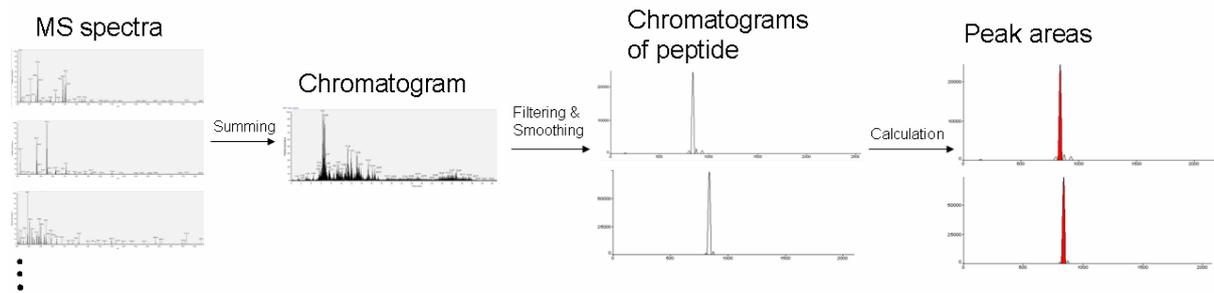


Figure 1: Schematic overview of the relative quantification process. The sum of all of the intensities of MS spectra (not MS/MS) results in an in-silico chromatogram which depicts the flow of masses into the mass spectrometer. For the calculation of a peptide only the mass flow of the peptide is of interest. Therefore contributions of the mass of the peptide are taken into account. The resulting chromatogram is smoothed afterwards. Due to the fact that the peptide can occur at different charge states several chromatograms have to be taken into consideration. The area below the chromatogram can be calculated as an indicator for the amount of peptide which entered the mass spectrometer.

2.2 Standards in proteomics

In recent years many proteomics technologies evolved. On the one hand they have driven this field while on the other hand this progress has led to a lack of comparability of proteomics studies. Proteomics researchers became aware that a minimum of description should be provided for experiments [68] to compare them effectively: a standard. The intention is to maximize the reusability and reproducibility of the experiments, and to easily disseminate data in an appropriate manner.

The first attempt to store proteomics data in standardized manner has been made by a collaboration of British institutes [44] in 2003. The Proteome Experimental Data Repository (PEDRo) [69] was published. The idea was to create a public repository, where all the proteomics data published should be submitted. The data stored was divided into 4 logical units according their activities:

- **Sample Generation:** This section stores data concerning the general acquisition of biological material, and general information concerning the design of the conducted experiment.
- **Sample Processing:** Here data concerning the processing of the biological samples is stored, including the techniques and the instrumentational parameters.
- **Mass spectrometry:** This section captures data concerning machine settings and machine types of data analysis by mass spectrometry.
- **MS results analysis:** Here the outcomes of mass spectrometry experiments are persisted, like peak lists and the results of search engines. Furthermore all the results of the analysis should be captured there as well.

The HUMAN Proteome Organization (HUPO) [33], which conducts many of the proteomics large-scale experiments, realized the need for action in this field as well. They founded the Proteomics Standards

Initiative (PSI) [34,35]. PSI set up several working groups covering the different areas needed for describing proteomics experiments. The domains set up by PSI comply largely with them described for PEDRo. In contradiction to PEDRo, the sample processing was subdivided due to the adherent and growing complexity of this section. The PEDRo schema strongly influenced the results achieved by PSI, since a lot of people involved in PEDRo are now members of the PSI. Before the proteomics community realized the necessity for standards, a similar standardization process could be observed in transcriptomics. They managed to establish the widely accepted Minimum Information About a Microarray Experiment (MIAME) [30,70] standard. According to this, the goal of PSI is to establish a standard similar in effect like MIAME for transcriptomics. Therefore, they named the newly evolving standard the Minimum Information About a Proteomics Experiment (MIAPE). Regular meetings of the PSI lead to a gradual evolution of these standards [36-43,71,72]. Due to the fact that the MIAPE standards are evolving, implementations must be more flexible and easier adaptable like it is demanded for databases complying with MIAME. Some of the guidelines still only consist of working drafts and others are just in planning phase. The currently available working drafts of the MIAPE standard are [73]:

- MIAPE GE v1.0 (Gel Electrophoresis – does not include image informatics): covers all the sections concerning gel electrophoresis in detail. “These reporting guidelines cover gel manufacture, and preparation, running conditions, visualization techniques such as staining, the method of image capture and a technical description of the image obtained. They do not explicitly cover sample preparation, but do require the recording of which samples were loaded onto a gel. They do not include spot detection or other analyses of gel images, nor do they include protein identification procedures” [74].
- MIAPE MSI v0.7 (Mass Spec Informatics): This section is similar to the MS results analysis section of the PEDRo. It stores the information required for protein and peptide identifications. “These guidelines cover the use of protein and peptide identification and characterisation software and the data generated. They do not cover the mass spectrometry that generated the data, or the reduction of ‘raw’ profile data to peak list” [75].
- MIAPE MS v2.1 (Mass Spectrometry): This module is similar to the Mass Spectrometry section of the PEDRo. It captures data concerning the mass spectrometry machine. “These guidelines cover both the operation of a mass spectrometer and the generation of mass spectra from the ‘raw’ data. They do not cover the delivery of sample to the mass spectrometer, or the interpretation of spectra by search engines” [76].
- MIAPE GI v0.1 (Gel [image] Informatics – online wiki draft for comment): This part covers the processing of the found proteins from gels. It includes spot detection, shape and intensity of the found items, and other information adherent to the analysis of found gel items. It consists just of a draft version where comments can be posted.

- MIAPE CC v0.2 (Column Chromatography): It encompasses separation of samples by column chromatography. “These reporting guidelines cover a column chromatography experiment from the selection and configuration of a column, through the selection of a suitable mobile phase, to the collection of fractions and associated detector readings. These guidelines do not explicitly cover a sample preparation procedure but facilitate the description of the sample. They do not include protein identification procedures. Where multidimensional chromatography is used, the material below is repeated as required for each dimension, with specific fractions from one column being used as the sample for another” [77].
- MIAPE CE v0.4 (Capillary Electrophoresis): This module describes the parameters of capillary electrophoresis experiments [78].

Furthermore, there is another module announced for the sample preparation and handling (MIAPE SP v0.1). Currently there is no information available about this module.

Additionally to the previously described PEDRo database another mentionable repository emerged from the European Bioinformatics Institute (EBI) [79]: The PRoteomics IDentification database (PRIDE) [80,81]. The idea was to provide a public repository for protein and peptide identifications together with the evidence supporting these identifications. Nowadays PRIDE is complying largely with the standards provided by PSI, which elevates and enhances proteomics research data made publicly available.

2.3 Software Technology

Java 2 Enterprise Edition (J2EE)

Enterprise applications are applications providing services to a wide range of users. The users expect from such a system reliability, scalability, high availability and security. “The Java 2 Enterprise Edition platform provides a component-based approach to the design, assembly, and development of enterprise applications” [82].

The J2EE platform provides a distributed component-based application model. A J2EE application is split in several tiers, offering the following functionality (see figure 2):

- Client-tier: is responsible for the presentation and the user-interactions normally comprised of a client application, an applet or a web-browser; it is executed on the client machine.
- Web-tier: handles communication between the web-client and the business logic of the server; implements the logic for the graphical representation to the browser; provides services for applets and client applications via hyper text transfer protocol (HTTP) and HTTP Secure (HTTPS); the components are executed on a J2EE server.

- Business-tier: represents the business logic of the application; comprises the execution of the logical procedures of the system; this encompasses calculations, preparation of data which has to be stored in the database, preparation of data for visualization, etc; it is located on a J2EE server.
- Enterprise-Information-System (EIS) –tier: this tier keeps the data and takes care of the persistent storage; it is located on a database server.

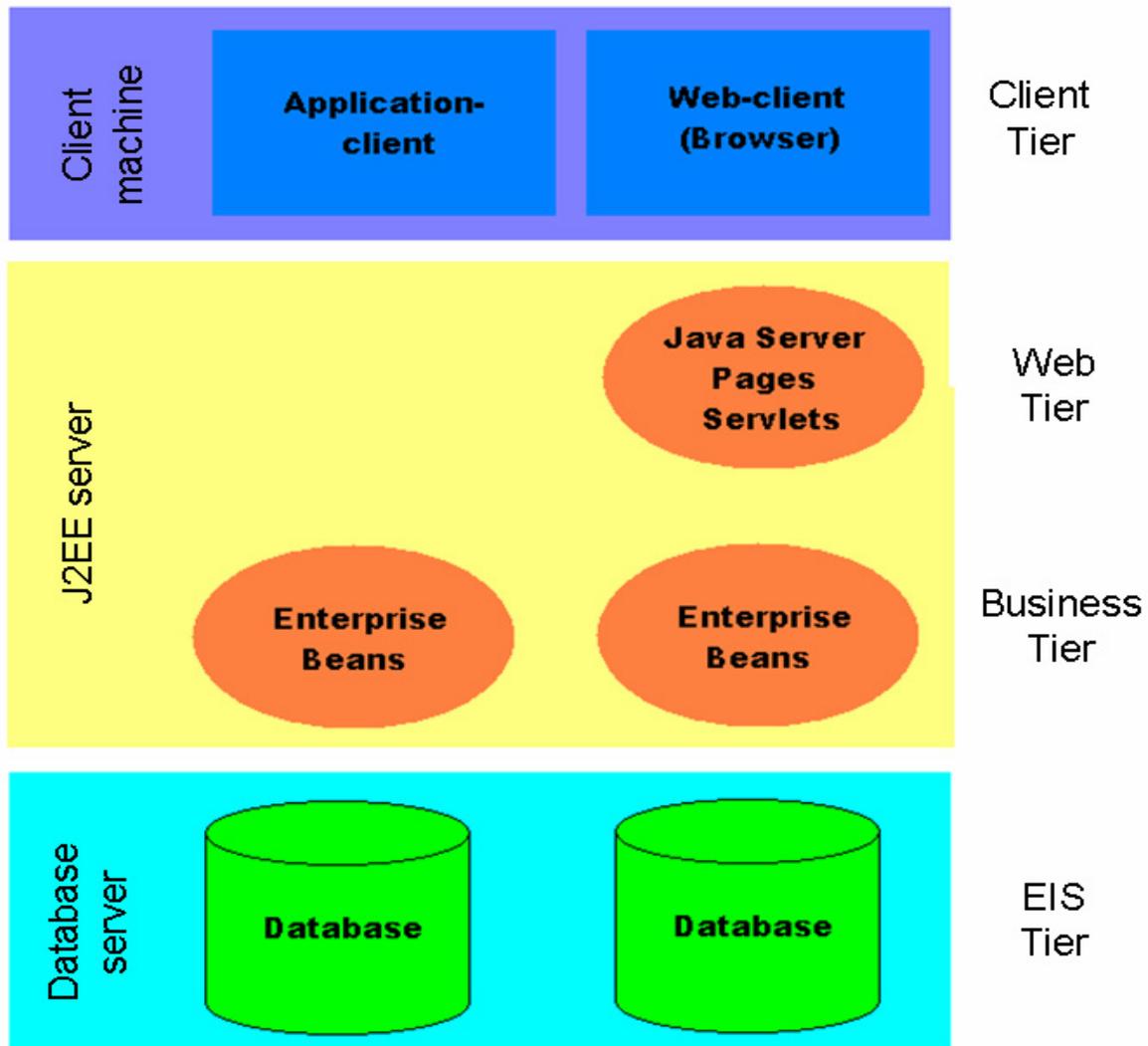


Figure 2: Schema of a multi-tiered application. The application is divided into 3 or 4 tiers respectively. Due to the separation of presentation-, business-, and data logic, any of the components can be upgraded or exchanged.

Since the web-tier and the business-tier can be located on the same J2EE-server or no web-tier at all is needed this component based model is entitled as 3-tier architecture. Due to the separation of presentation-, business-, and data logic, the components can be changed, upgraded and exchanged. This poses a major improvement regarding the maintainability of the software system compared to monolithic systems, and introduces independence from specific technology implementations (different

relational database management systems (RDBMSs) are supported; different J2EE compliant servers can be used; etc).

Business-tier – Enterprise Java Beans (EJBs)

Enterprise Java Beans (EJBs) can be used to implement the business-tier of J2EE applications. An EJB is a server-side, application-server-independent software component, whereas a component is a reusable piece of software. They can be developed and assembled easily to create sophisticated applications with high efficiency [83]. They are running in a so-called EJB container. Containers are the interface between components and low-level platform-specific functionalities. EJBs allow the development and maintenance of large-distributed server applications [82]:

- The EJB container manages the system-level tasks like the transaction-management, the authorization, etc. The developers can focus their work on coding business solutions regardless of low level programming tasks.
- The business logic is encapsulated by EJBs and the client programmer's work is limited to the client-side presentation, neglecting the coding of business rules or database transactions. As a benefit the client-side applications do not need to run on high-end performance machines.
- EJBs are portable and reusable components. They can be used by multiple applications.

Generally, there are three types of beans:

- Session-Beans: execute business tasks for the client
- Entity-Beans: represent Java objects persisted in a database
- Message-Driven-Beans: are similar to session-beans, but they are used for the asynchronous execution of business tasks without user interactions

Session Beans are used to implement the business logic. They provide business methods for client applications or for the web-tier and communicate with entity beans to retrieve relevant information from the database. On this level the processing of information should be performed e.g. the execution of algorithms, or the assembling of information for the presentation-tier. According to their behavior two types of session beans are differentiated [82]:

- Stateless session beans: correspond to a single request and response conversation. A client requests a task execution from the bean (execution of business code) and receives a response. When the transaction is finished the session bean is released and the request specific information during the task execution is discarded. The advantage of stateless session beans is

the reduction of creation overhead at method invocation. Therefore multiple pooled stateless session beans can act simultaneously.

- **Stateful session beans:** one business processes can span more than one single HTTP request/response cycle. Several requests are necessary to achieve the desired outcome. This corresponds to a multistage conversation between the server and the client. In case of a stateless session bean the necessary information has to be fetched several times from the database. However, stateful session beans persist data in the session context and can offer it several times without repeated database access. This proves to be extremely useful when for example the same results of a time-consuming data transaction are needed for several requests. The same results are available to the same client for several requests. The data will be lost when the session expires or the browser is closed. A disadvantage is the higher memory consumption of stateful session compared to stateless ones.

The choice of the appropriate type of session bean depends on the type of task which has to be performed by the business logic.

Entity Beans represent entries in databases. Normally, one Entity Bean corresponds to a table entry in a database with the attributes adherent to it. They implement an object-oriented approach of accessing the data stored in the database and provide methods for entering and deleting persisted data. The attributes of the objects are accessed by getter and setter methods. The persistence of data to the database via entity beans can be achieved by two approaches [84]:

- **Bean managed persistence (BMP):** the access code to the database has to be written by the developer. The developer is responsible for the synchronization of data with the underlying storage engine. Additionally transactions and locks on the data are fully managed by the application code. This approach provides the full control over all data storage actions including database specific optimizations.
- **Container managed persistence (CMP):** the data access and storage is handled by the EJB container. The developer does not need to take care of low-level actions and can concentrate on the development of business logic. Some containers provide caching and read ahead mechanism which can be more performant than the custom application code of bean managed persistent entity beans.

Message Driven Beans are a special peculiarity of session beans. They consume messages and are invoked asynchronously. The message listener who initializes the beans is listening to a so-called Java Messaging Service (JMS) topic/queue, to which messages from the client are sent. When the message driven bean consumes a message from the queue the working process starts. As soon as the bean

finishes its work the initiating message is removed. Message driven beans can invoke during bean execution additional session beans, entity beans, or retrieve information from any other resources needed (e.g. a file system). The benefit is that the client is released after committing the message and does not need to wait until its task is finished. This asynchronous strategy proves to be quite useful when long lasting processes have to be started. Furthermore, messages can be consumed concurrently, allowing parallel-processing of tasks [85]. When the tasks are computationally intensive, the maximum number of concurrent jobs can be set to a fixed number avoiding server overload.

Session façade design pattern

A pattern is a proven solution for a software design problem which occurs over and over again. They are reusable solutions reflecting the experience of programmers coming across similar problems [86]. One of these patterns is the session façade design pattern, which has been extensively used throughout this thesis. This pattern hides the complexity of the object-interactions of the business logic from the client and minimizes the time-consuming interactions between client and server. Session beans should encapsulate the business logic of an application. This pattern guarantees a minimization of parameters send to, and received from the client, but the complexity of the server-side components is increased (see figure 3). Whereas the complex business logic is executed in session-beans, while the storage of the data is matter of the entity beans.

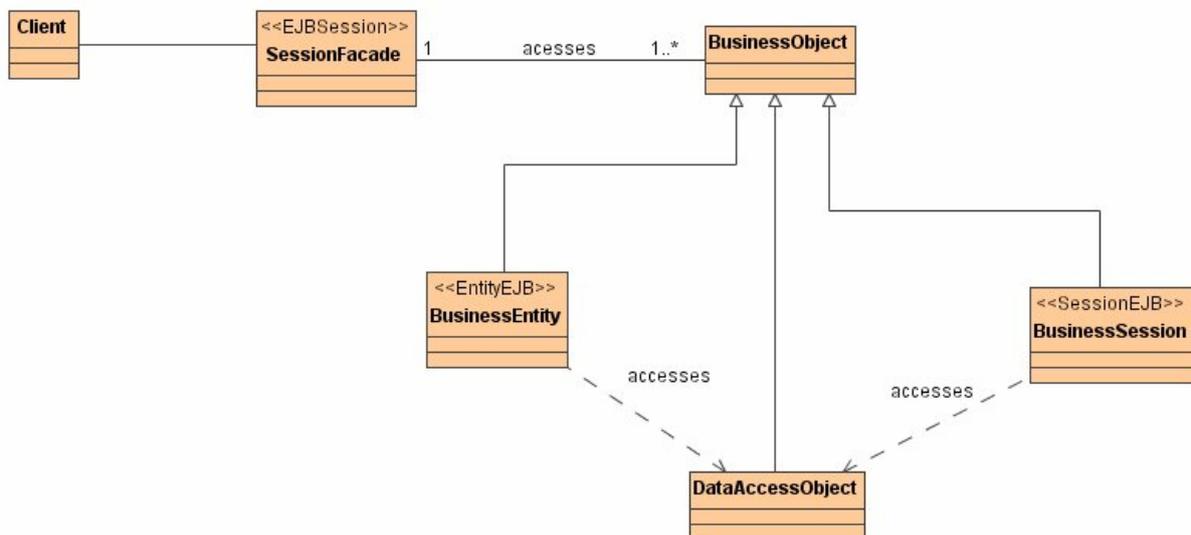


Figure 3: Class diagram of the session façade design pattern [87]. The client communicates exclusively with the session façade, while the façade accesses several business objects, like entity beans, session beans or data access objects. The entity of access operations is carried out over one session bean.

The benefits of this pattern are [87]:

- The couplings between client and business objects are loose.
- Direct interactions between client and server are minimized. As a matter of fact the networking overhead is reduced.
- A uniform interface is exposed. This abstraction simplifies the interfaces and it is easier to discover the relevant methods.
- The amount of business-objects on the network service layer is reduced (the layer where the client has remote access).
- The workflow is centralized, hence the underlying interactions and dependencies are hidden behind the session façade.

Session façade patterns use one or more data access objects (DAOs – common Java objects) to persist the business data. Calculations and other modifications of the data are performed with common Java objects. Consequently the entities stay persistent and they are not affected by business operations.

Web-Tier – Java Server Pages, Servlets

Java Server Pages (JSPs) and servlets form the web-tier of the developed application. In contrast to the business-tier they do not focus on the solution of business problems, but represent one possible presentation layer within the J2EE platform.

Servlets are web-enabled Java classes creating dynamic web content. They are waiting for HTTP requests and return HTTP responses. The programmer has to code the section from the request to the response, while the underlying programming details for the request/response conversation stay hidden. They can be compared to common interfaces hosted by a web-server, generating dynamic web content [88].

JSPs are servlets, which are dynamically generated. They utilize the same request and response mechanism for the generation of dynamic web content, but the creation approach is different. The JSPs are not written in plain Java code. They are using syntax similar to hyper text markup language (HTML) in respect to the tagged structure. However behind a JSP tag a Java class with defined interfaces is performing the HTML rendering. In contrast to servlets, HTML code can be directly integrated into JSPs, which eases the programming. The JSP template files are compiled at run-time to Java servlet classes.

The combination of both technologies enables highly efficient web-development. While JSPs are suited more for dynamic HTML pages, servlets provide considerable advantages concerning programmatic intensive approaches, like retrieving information from EJBs or rendering of images [88]. Similar to EJBs, JSPs and servlets must be hosted by a container (here a servlet container).

Jakarta Struts

The J2EE platform has shown that the compliance to standards implicates reasonable advantages concerning maintainability and easier development of components. For the development of web-applications using JSPs and servlets there are two general models [89]:

- Model 1: the web-application is comprised of an accumulation of JSPs, which work independently from one another. The JSP alone processes the requests and responses to the client (see figure 4).
- Model 2: is based on the model-view-controller (MVC) principle. It consists of a model which encapsulates the business logic, the view which corresponds to the JSPs and a set of servlets to guarantee centralized process-management (see figure 5).

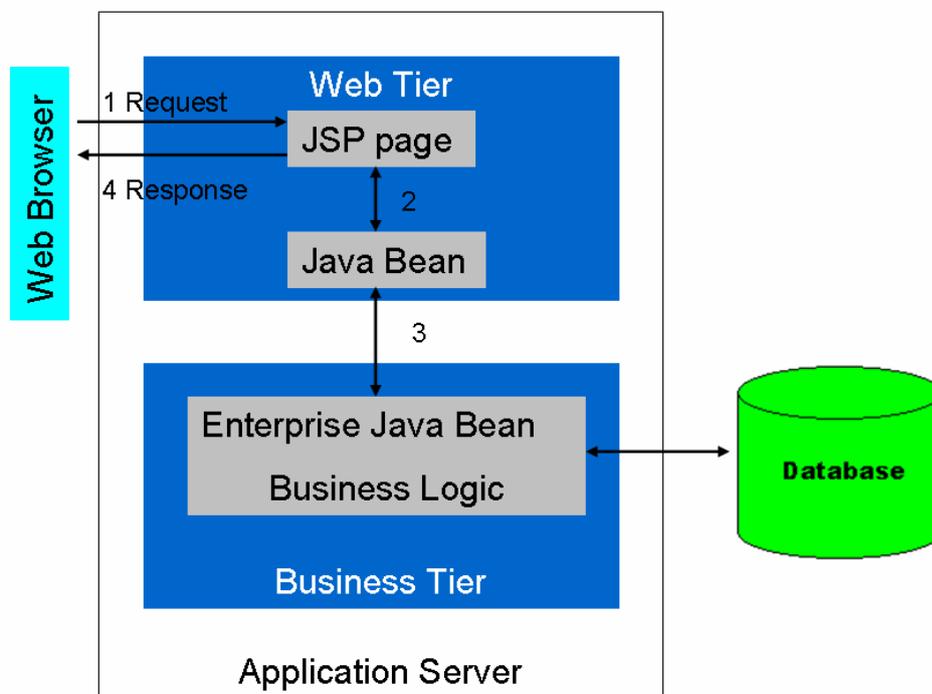


Figure 4: JSP model 1 architecture. The JSP alone processes the requests and generates the responses to the client. The JSP page receives a request from the browser, and communicates according to the type of the request with an EJB. The EJB is talking to the business logic which is connected to the database. The JSP page replies the request.

In Model 2 all of the occurring events have to be processed by the controller. The controller is a servlet that decides if an event is valid for the view, and informs the view about changes. The new information is evaluated by the model. If there are any changes, the view fetches the new data and depicts them. Therefore the responsibilities for request handling are split into three domains:

- Model: encapsulates application state; responds to application state questions; exposes application functionality.

- View: renders the model; is updated by the model; presents the data to the browser.
- Controller: defines the application behavior; manages user actions to update the model, selects the corresponding view for the response.

The advantages of the MVC approach are:

- Segregation of the model from the view; only a defined interface is necessary.
- Several views can access one model; there must be ascertained that the controller notifies every view if there is a change in the model.

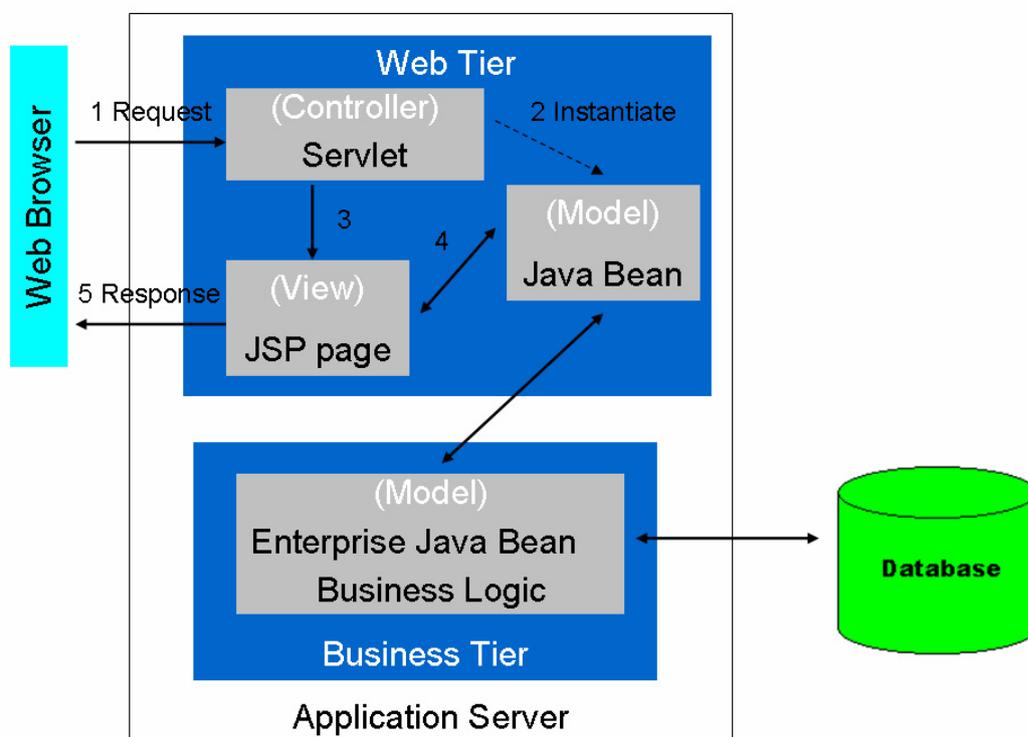


Figure 5: JSP model 2 architecture. The controller receives the request and creates any beans or objects needed by the JSP, and forwards the request. In detail: The controller receives the request from the browser, instantiates a bean which is connected to the business logic of the server. The controller communicates with the view. The view communicates with the bean. The JSP replies to the browser.

The major advantage of model 2 is the better maintainability and extensibility of the system. Therefore, a model 2 approach should be chosen for large web-applications, for the purpose of prototyping it is better to choose a model 1 approach.

The Jakarta Struts project provides a reliable and tested open-source framework for the development of J2EE web applications using JSP and servlet technology. The MVC concept has been fully implemented by the Struts framework providing all the advantages mentioned before. A lot of commonly used JSP-tags are already provided by the framework. Additionally the easy extensibility

with self-written tags makes Struts to a powerful framework for the development of J2EE web applications.

Model Driven Architecture

Although J2EE provides an environment well suited for the development and maintenance of server application projects, the implementation of required interfaces for EJBs and configuration files for Struts, etc result often in programming overhead which slows down the development. Nevertheless, the prolongating work consists of tasks, which can be performed automatically. To overcome these shortcomings and to concentrate on the implementation of business logic a model driven approach for the implementation of the platform has been chosen.

These handicaps of general development platforms has been recognized by the Object Management Group (OMG) [90], which was formed to reduce the costs and the complexity of new software applications. The OMG introduced the Model Driven Architecture (MDA), whereas a model is the “formal specification of the function, structure and/or behavior of the system” [91]. “The MDA addresses integration and interoperability spanning the life cycle of a system from modeling and design, to component construction, assembly, integration, deployment, management and evolution” [92]. To achieve this, different technologies had to be applied, including the XML Metadata Interchange (XMI) format, Unified Modeling Language (UML), and Meta Objects Facility (MOF) [91]. The idea behind MDA is to depict the conceptual architecture of an application in a model. The model is afterwards transformed into models of different abstraction level or to source code. The layers of abstraction for a model can be divided into platform independent models (PIMs) and platform specific models (PSMs) [93], whereas platforms in this context means a platform like J2EE [94] or .NET [95]. The PIM is applicable to multiple platforms. The transformation works from the higher levels of abstraction to next lower level (see figure 6).

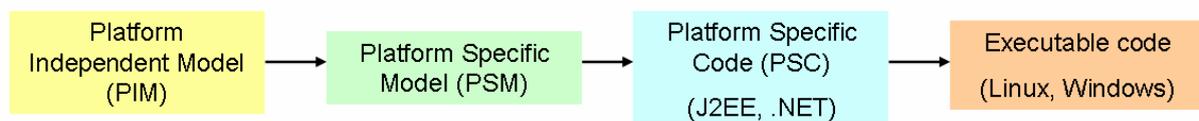


Figure 6: Workflow of model driven architecture. The platform independent model has to be transformed to a platform specific model; which has to be transformed to the platform specific code; the code has to be transformed to executable code (compilation).

Defined rules should be compiled for the transformation from one layer of abstraction to the next one. Actually, the research of the OMG goes in the direction of elaborating standardized rules for the transformations [96-98], because the major benefit of such an approach emerges when the transformations work in an automated manner.

AndroMDA

AndroMDA is an open source code generator [99]. A generator is a program which takes models at a higher level of abstraction and transforms it into the next lower level. A code generator creates platform specific implementation code (e.g. Java, C++, C#, etc). But how to describe software models at PIM state? One solution is the OMG standard unified modeling language (UML) which is a language for describing software applications. For the visualization of the UML model many Computer Aided Software Engineering (CASE) tools like MagicDraw [100] or Poseidon [101] are available. The CASE tools can store the UML information either in proprietary formats, which can be read only by internal code generators, or in XML Metadata Interchange (XMI) format. The stored file is afterwards read from the code generator and the UML model is interpreted according to the corresponding code generation rules and transformed into implementation code.

AndroMDA is able to read the UML model from the XMI format and makes mainly use of the UML class diagram for the software description. Like in Java programming language a class can correspond to a real Java class or to a more abstract set of classes and interfaces like an EJB. For the rough classification of the purpose of the used classes, so called stereotypes are assigned to them. E.g. the stereotype <<Entity>> corresponds to the generation of an Entity Bean, while <<Session>> would result in the generation of a session bean. A class can have several attributes (like the columns of database tables), and/or several operations (like access methods, or the business methods of a session bean). For a more detailed description of the functionality, the model provides tagged values. Moreover AndroMDA features a lot of possibilities to map relations between classes. Stereotypes and code generation rules can be extended by the developers individually to their needs which make AndroMDA to a flexible and powerful tool. Nevertheless, AndroMDA is still largely independent from the targeted platform, since e.g. <<Entity>> could correspond to any type of classes that persist something in a database.

Figure 7 depicts a detailed scheme of the used build process for the application development. In a first step the application is modeled with MagicDraw [100] and stored in XMI format. The whole build-process itself is determined by an adjustable Ant [102] build file. There the sequence of the invoked steps is defined. First the XMI and the code generation rules written in velocity [103] are read and the syntax is checked by the AndroMDA code generator. The cartridges contain the meta-information how to transform the information stored as stereotypes, tagged values, etc in platform specific code. The mapping of cartridges to stereotypes is definable in the file andromda-cartridge.xml. The andromda-cartridge.xml is exhibiting so called outlets which can be executed by the AndroMDA ant-task. Therefore every single step of the code-generation process is controllable and the cartridges are flexible to be adapted by the developers to their demands. When the model and the syntax had been validated successfully the AndroMDA code generation starts.

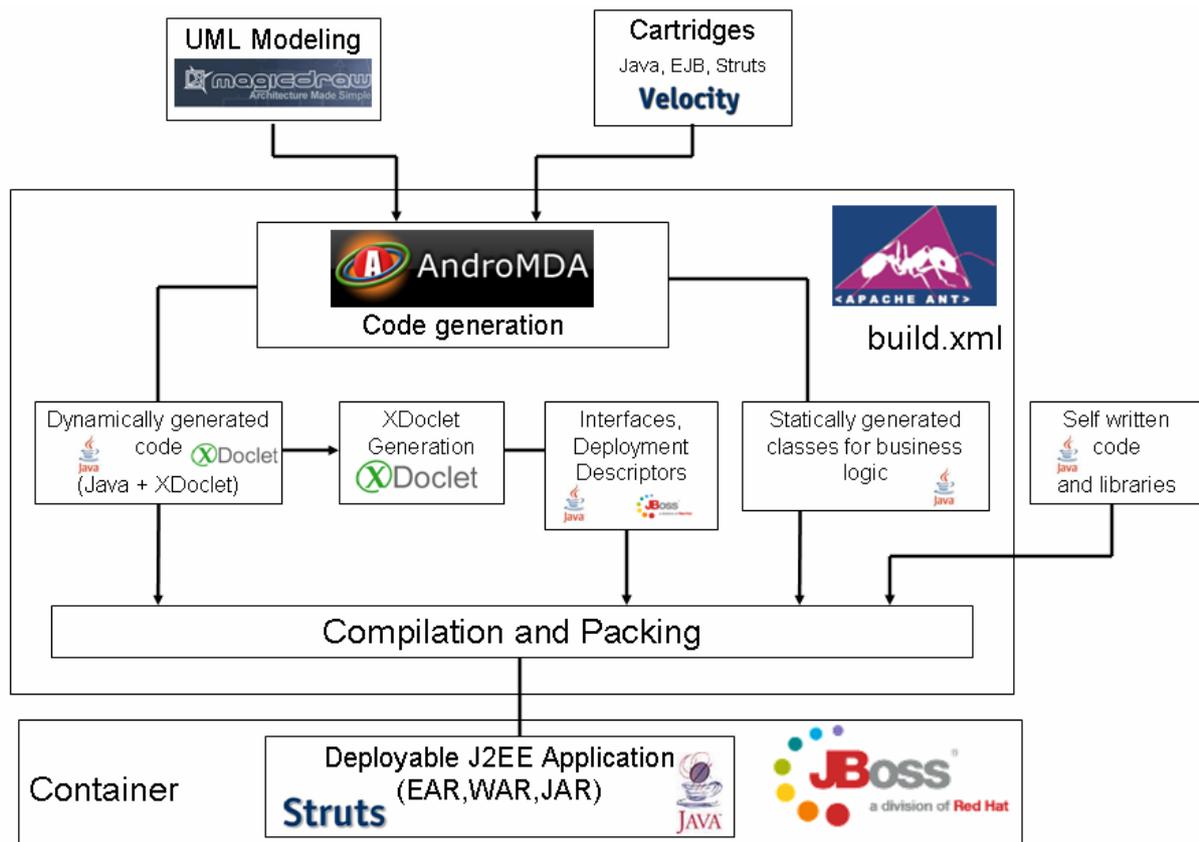


Figure 7: Overview of the AndroMDA build process used for the development of a J2EE application. The model and the code generation rules are provided to AndroMDA and J2EE specific code is generated. The building process covers all the steps from the interpretation of the UML-model over the code generation to the compilation and packing of a fully deployable enterprise application.

The output is split in two groups:

- **Dynamically generated code:** The code is overwritten every time the AndroMDA code generation process is called. A latter change in the model induces a direct change in the code. This proved to be extremely useful in the generation of entity beans, or interfaces provided by session beans, or DAOs. Subsequent changes and extensions in the data model are easily manageable.
- **Statically generated code:** This code is generated only when a class is defined the first time and is never overwritten unless the physically generated file/files is/are deleted. This is applicable for code that cannot be generated automatically. This code can be seen as a template and must be extended to implement the business logic. An example is business code. The EJB interfaces and classes are generated automatically, but additionally implementation classes are generated which extend the automatically generated ones. The regeneration of such classes would result in loss of business code. A further example would be the generation of Struts code for the presentation-tier. It is reasonable to generate input fields, display fields, the standard configuration entries for Struts or standard functionality like storing and retrieving of data automatically, but only in the manner of templates. It makes no sense to lump together

the display for all database applications. There are always specific problems to solve, where the code has to be adapted. If there are changes in the automatically generated code that affect the statically generated part the changes have to be adapted manually.

The combination of these two types resulted in an efficient and still flexible approach for the development of large enterprise applications [104].

The automatically generated code contains XDoclet [105] tags. After the AndroMDA code generation has been finished the XDoclet code generation engine uses these tags to generate the local, remote and home interfaces for the EJBs. The deployment descriptor files are generated as well, to enable the compatibility to a JBoss Server [106]. In contradiction to the AndroMDA code generator XDoclet is not platform independent but is just applicable to Java. Nevertheless, it is not restricted to JBoss, and supports several J2EE application servers. Afterwards the whole code is compiled and packed to the corresponding deployable files (EAR, JAR, and WAR).

Relational database management system (RDBMS)

The reliable storage of data from J2EE applications is guaranteed by sophisticated relational database management systems [31,107,108]. A relational database model consists of a series of unordered tables, connected with one another. The design of a database should be a model of the real world. The tables represent the items (entities) of the real world. The tables are composed of rows and columns, where the uniqueness of the rows must be guaranteed by primary keys. The relations to other tables are established by the export of primary keys to other tables (foreign keys). This principle allows the mapping of one-to-one, one-to-many, and many-to-one relations between real world objects. For many-to-many relations an additional helper table must be added. Codd [109] elaborated rules to normalize data for storage in a flexible, non-redundant and efficient way. Many of the major relational database management systems [110-113] comply with these rules.

The operations on database tables are normally performed by the execution of Structured Query Language (SQL) statements. Although, SQL should pose a standardized language for database access, each of the databases uses a kind of specific dialect, and the syntaxes differ within the RDBMS implementations. E.g. the statements for fetching data are mostly the same, while statements for creating tables vary largely, due to the use of different data types and the different organization of data storage. An additional reason for differences in SQL implementations is that RDBMSs and their features were developed independently. Nevertheless the entity beans and the EJB-QL (EJB Query Language) [114] of the J2EE platform take over the burden of programming RDBMS specific SQL statements. However the use of EJB-QL takes longer than the direct execution of SQL statements. Therefore this approach is more feasible for handling smaller datasets having the advantages of better maintainability. Furthermore it is less flexible because every type of query in EJB-QL is hard-coded. Therefore a flexible combination of user-specific constraints for queries is not easy to implement. In

response to this drawback the Java Database Connectivity (JDBC) API can be directly utilized [115]. JDBC is a standardized programmatic interface for Java applications, which provides RDBMS corresponding drivers. It is a low-level application interface (API), since it executes SQL statements directly. It has been developed as fundament for interfaces operating at higher level, offering a more user-friendly API. JDBC performs the following 3 steps:

- Establishes a connection to the database.
- Sends SQL statements
- Fetches the results

JClusterService

The amount and the complexity of data for mass spectrometry experiments increase dramatically. The algorithms handling the vast amount of data pose computationally intensive tasks slowing down the work progress of users accessing the data. Therefore, outsourcing and parallel-computing is a requirement to reduce overall analysis time.

The JClusterService is a web service for delegating resource intensive tasks to a high performance computing cluster system [116]. The flexible modular design allows the integration of any command-line tool. The tasks are executed on dedicated computational nodes and do not slow down the main web-server anymore. JClusterService is a J2EE application, providing a remote Java API accessible by Simple Object Access Protocol (SOAP) [117].

Authentication and authorization

A system located in a multi-user environment must meet certain expectations:

- Users must be authenticated.
- The system must be secure.
- The roles for data manipulation capabilities must be definable.
- The user should have the possibility to share data with other users of the system.

The in-house developed Authentication and Authorization System (AAS) [31,118] offers an easy-to-integrate environment for the administration of multiple applications and multiple users. On the one hand it provides an API for Java applications, on the other hand it offers specific JSP tags for Struts web-applications. For authentication a combination of user name and password is required. The data manipulation roles can be defined via fine-grained Access Control Lists (ACLs). The ACLs are assignable to groups of users or users individually. Additionally the system offers the administration of institutions, and control features like logging of user accesses.

2.4 Experimental procedures

In order to validate the system under experimental conditions a biological experiment was conducted. The proteins serumalbumin [GenBank:AAA51411.1], human apotransferrin [ref:NP_001054.1] and rabbit phosphorylase b [PDB:8GPB] were used. Protein aliquots were split to two samples. To label the proteins ICPL (isotope coded protein label) was used. One aliquot was labeled with the ^{12}C isotope (light), the other one with the ^{13}C isotope (heavy). From these stock solutions samples for MS/MS analysis which contained defined ratios of heavy and light isotopes were prepared by mixing the solutions of light and heavy labeled peptides. To separate peptide mixtures prior to MS analysis, nano reverse phase high-performance liquid chromatography (nanoRP-HPLC) was applied. 500fM of each mixture was separated three times using the same trapping and separation column to reduce the quantification variability error which is introduced by HPLC and mass spectrometry. The exit of the HPLC was online coupled to the electrospray source of the mass spectrometer. PFF database searching was done with the Mascot Daemon [11] using an in house database of the Institute of Molecular Pathology in Vienna [119] (for a detailed description of the experiment see the attached paper: MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data). The results of the database search and the raw data were used for the test of the developed proteomics platform (see 3.4).

Chapter 3

Results

3.1 Extensions to the AndroMDA environment

The proteomics standard MIAPE is still evolving. MIAPE compliant database implementations must therefore be flexible and easily adaptable. In view of this we chose the model driven approach using AndroMDA. Despite its utility, it was however nevertheless necessary to extend the AndroMDA code generator described in 2.3 to enable rapid development.

A major problem was posed by the generated web-interfaces which displayed the entered data only as long lists. A database application should provide more. The user should be equipped with a set of tools to organize data, or better still to filter out irrelevant information. At the beginning of the project a ready-made solution to this problem was available from Struts: the display tag library. This provided tags to easily display tables with a definable number of elements per page, the ability to turn pages, and to sort by the columns of the table. Although in principle a good approach it however suffers from some drawbacks compared to the solution described in this thesis:

- 1) The displayable columns cannot be faded out individually and the page is thus not customizable for the user; this feature is especially important when many columns must be displayed; the user can easily lose the overview.
- 2) The tables provided have only a display purpose, i.e. there is no interaction with the EJBs or any other possibility to post queries. Such features must be implemented separately and manually. Such an approach completely contradicts the paradigm of model driven architecture.
- 3) Worst of all, all of the entries must be retrieved from the database and loaded into the web session, i.e. selective retrieval of only the items needed for the displayed page is not possible. The consequence is a significant loss of performance with large datasets and unjustifiable response times from the system.

Usefully the MARS application, developed in-house for DNA microarray studies [31] provided a ready-made solution to the problems described above, the so-called “report bean” [120]. Direct coupling to the database is achieved via a session bean executing direct JDBC statements and entity beans for fetching the Java objects. The JDBC statements are assembled by a generic JDBC generator relying on abstract Java objects representing database columns, with the columns for one bean possibly originating from different database tables. This generic approach allows for great flexibility in the composition of a variety of possible query fields without requiring any SQL knowledge from the end user. The high efficiency of this approach stems from the strict separation of the querying and fetching phases. During the querying phase only a minimum amount of information (normally only the primary keys) is fetched from the database. The information in other fields is not needed since it is provided in the query itself. For display purposes the entity beans only fetch information for the items to be displayed which, depending on the chosen page size, results in 15 to 100 elements rather than the 20000 that can occur in the result lists for proteomics experiments. The bean implements an interface where standard methods needed to display the results are defined. As well as querying and fetching elements, the bean possesses functionality that includes determining the number of elements and pages, turning to the next, previous, or a distinct page, or retrieving the queryable fields and corresponding operators (‘=’, ‘>’, ‘LIKE’, etc).

The implementation of report beans however entails much time-consuming and error-prone work due to their inherent complexity. Report beans have therefore been integrated into our in-house developed extension of the AndroMDA project [121]. The UML-based implementation is very simple (see figure 8). The session bean `EntityBReportService` corresponds to the report bean. The only difference to a normal session bean is the tag `@andromda.service.reportbean`. This must have a reference to the corresponding entity bean in order to populate the queryable fields. Additionally the entity bean must be associated to a constants class (`GlobalConstants`), which generates constants for the persistent fields which can be used by other classes. The `SharedEntityB` corresponds to a sharing table, which is needed only when sharing mechanisms (explained later in this section) are desired. Extensions to the generated report beans (like queries to other database table fields) can be implemented gradually. The generated classes nevertheless form a reliable basis for further development.

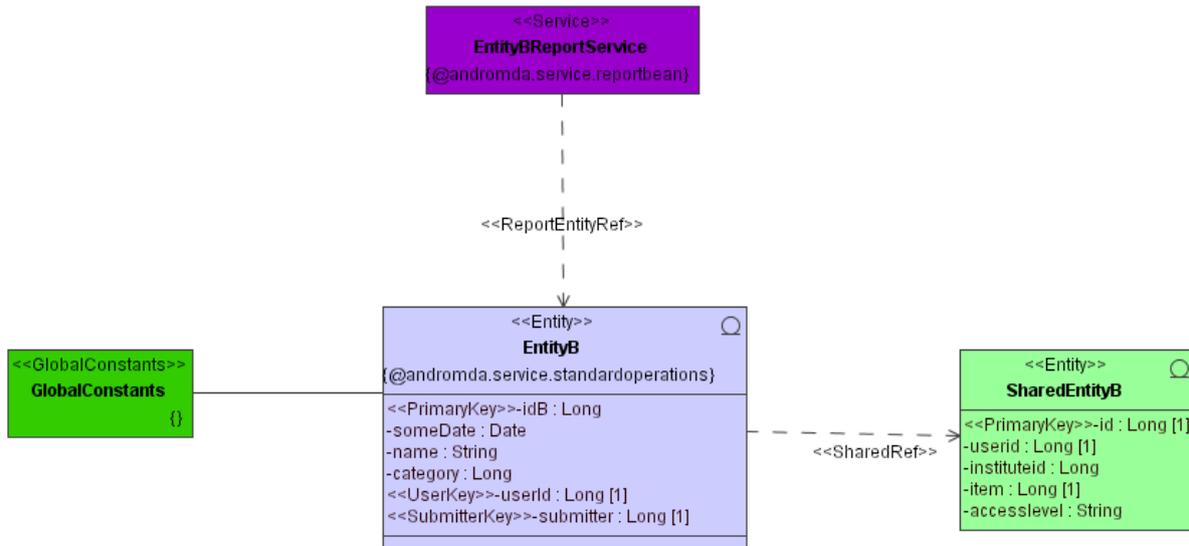


Figure 8: UML-model of a report bean for AndroMDA. `<<Entity>>` corresponds to an entity bean. `<<GlobalConstants>>` is a class where constants for other classes are generated and can be used. The `<<Service>>` with the tag `@andromda.service.reportbean` is the report bean.

Although report beans form the business logic for the execution of sophisticated table operations, they are not the web-interfaces for the user. The Struts servlets and JSPs are still missing. The in-house developed MARS application [31] already provided solutions with common look and feel for Struts servlets and JSPs. These solutions did not however exhibit the clearly structured segmentation in functionality that is needed for code generation. The methods provided by the servlets were therefore structured according to their functionality. A number of methods were evolved including `createEdit` (handles everything concerning storage to the database), `changePageSize` (to enter a different number of items for one page), `submitQuery` (to execute a query via a report bean), `scrollNextPage`, and so on. Despite this clear structure, the development of pages using report beans remained a very time-consuming and error-prone task that was also much more error-prone than the development of report beans. Every step and method had to be kept in mind.

This technology was therefore integrated in the AndroMDA cartridges (see figure 9). This model can be viewed as an extension to the previously described report bean model (see figure 8). The main extension is the class `EntityBReport`, which is responsible for the generation of the Struts servlets, JSPs and forms (value objects which are sent via HTTP, whereas value objects are plain Java objects exhibiting getter and setter methods for their attributes). This must have a dependency to the corresponding report bean. The fields that must be displayed and additional information about them is retrieved from the association to the entity bean. This additionally generates standard field display names to facilitate internationalization of the application. Then it renders extensible entries for the Struts controller, which controls the workflow of the pages. Furthermore some additional features are definable. The association to the `WebConstants` is needed, since the references between several classes (and the lookup in the web) are handled by them. The dependency to the `TemplateServiceFactory` is

used for storing user input in the database. This part can however be used for all entities (since it is part of the standard AndroMDA implementation [121]).

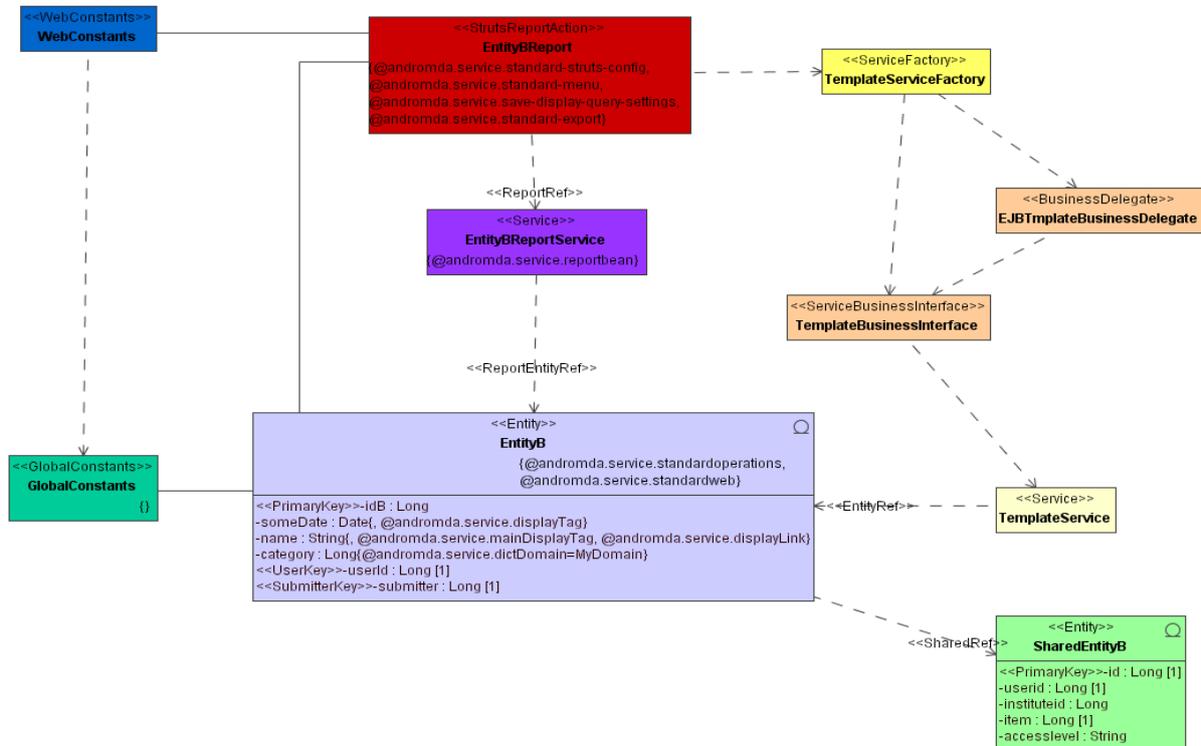


Figure 9: UML diagram for a fully functional report page, providing lists with query, sorting, and scrolling features, and input masks as well as the possibility to share with others, create edit and delete functionality. The model is an extension of the report bean model (see figure 8). The responsible class for the generation of the Struts servlets and JSPs is the <<StrutsReportAction>>.

The result generated by this model is a ready-to-use web-interface which provides multiple useful features including page scrolling; definable page size; sorting by all of the available columns; querying; saving of default queries; customization of display fields; saving of default display fields for each individual user; sharing of data with other users and institutes; controlled access rights to the data via the AAS [118]; data manipulations including the creation, editing, and deletion of data objects (see figure 10).

These generated pages can furthermore be adapted and changed to meet specific needs and therefore form an excellent basis for further development. The rate of development of highly-functional web-applications was moreover dramatically increased: the manual development of such a page took 3-4 days (even when routine); the same procedure can be completed with AndroMDA in 1-2 hours.

The figure displays two web pages from the AndromDA system. The left page, titled "Software", features a query builder at the top with fields for "Name" (containing "XC*"), "DateOfRelease" (containing "> 10Oct2006"), and "Role". Below the query builder is a table of software entries:

Nr.	Name	Version	DateOfRelease	Role	User			
1	XCalibur	2.0	2007-03-05	massSpectrometrySoftware	Juergen Hartler			
2	XCalibur	testSoftware	2006-11-29	massSpectrometrySoftware	Juergen Hartler			

The right page, titled "New Software", shows a form for creating a new software entry with fields for "Name" (containing "XCalibur"), "Version" (containing "1.0"), "DateOfRelease" (containing "06.03.2007"), and "Role" (containing "MS-Controlsoftware"). A "Create" button is located at the bottom of the form.

Figure 10: Two web-pages generated by AndromDA without any additional programming effort. The page provides the combination of any type of queries, and the individual saving of them. The display fields are customizable and storable to each user individually. Both of the boxes can be hidden. Below there is a scrollable and sortable list. The fields on the right side of the table entries have the following meaning (from left to right): edit, share, delete. The page on the right depicts the automatically generated input mask.

The AAS (see 2.3) provides an excellent basis for a standardized user administration. Standard features are offered for the administration including the granting of rights to groups or users for different kinds of applications [118]. Some possible security leaks have nevertheless been fixed during this thesis. For single-sign-on purposes between several systems using the same AAS a cookie must be used. In the former version the username and the authentication identifier issued by the AAS were stored there in plain text. If a user did not log out properly, the cookie remained in the browser. Although the authentication id expires regularly, the use of the same browser by an unauthorized person, or the copying of the login and the authentication identifier is possible until expiration. We therefore implemented a Diffie-Hellman key agreement [122] to retrieve a key for the encryption of the information stored in the cookie. The key is changed regularly and can only be retrieved by registered applications of the AAS. Additionally, the expiration time of the authentication identifiers can be set by the client applications individually. At log-in-time the AAS receives this information and stores it in the database. As long as the user is active the time is updated. When this time is exceeded the corresponding authentication identifier is automatically removed and not at fixed time points as before. The critical time for the abuse of the identifier is thus minimized. The received key of the Diffie-Hellman key agreement provided a further benefit. In the former version the private key of the application and the passwords were sent as plain text (although over an HTTPS connection). These keys are now additionally encrypted and the sections where encryption and decryption occur are encapsulated, further augmenting the security of the systems.

The system furthermore revealed a conceptual security leak. The AAS is designed for the management of several applications. The central management of all of the settings by a central administrator is not feasible since the number of applications is steadily growing. A separate administrator is therefore in

charge of the rights for each application. The administrators are only permitted to assign rights to their applications. Rights can however be assigned to groups to simplify the management of the system, the granting of rights to each user individually would be error-prone and unnecessarily labor intensive (although possible). User groups can be reused in several applications. An administrator of one application was previously able to add his/her own or other accounts to groups (even with a lot of permissions) of other applications. In the new version, the group is first assigned to an application. Only the administrator of this application can now permit the use of the group in another application. The rights concerning a given application are thus fully controlled by the respective administrator.

As well as supporting several applications, the AAS supports the administration of several independent institutes, since data management on an institute level frequently makes more sense than on an individual user level. Nevertheless, the old AAS permitted only the assignment of users to one institute. Users can now be added to several institutes and institutes can be organized hierarchically, i.e. the members of a parent organization have access to data produced by sub-organizations.

In the course of this thesis the AAS has been integrated into the AndroMDA environment [121]. All the features provided were integrated. The AAS Struts tags for the access permissions to the data are generated by default, and sharing mechanisms can be generated automatically if desired. The data can be shared with users and/or institutes (see figure 11). Only the owner – which can be a user or an institute - can share data objects and grant read, delete, and update permissions. The code can be rapidly generated and fine-grained access control provided.

The original AndroMDA version was not designed to work out-of-the-box with different RDBMSs since the primary key generator used was in fact a random key generator. An issued key could thus be reissued. MARS [31] already provided a functional key generator for Oracle, which has been integrated in the AndroMDA project. According to this additional key generators for PostgreSQL and MySQL have been implemented. The application detects which database is used and executes database specific statements. To modify RDBMS it is only necessary to replace the corresponding JDBC-library and the database deployment file (containing access information including address, username, and password).

Further to the above, a generic file upload zone, extracted from MARS, where even large files can be uploaded (tested up to 2 GB).has been added to the extended AndroMDA project.

Sharing



You are about to share item: **testCentroid**

	Name	E-Mail			
<input type="checkbox"/>	Bioinformatics Group	zlatko.trajanoski@tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Institute of Pathology, University of Graz	karin.wagner@klinikum-graz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Inserm U255	jerome@irgendwas.fr		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Visitors	none		<input type="checkbox"/>	<input checked="" type="checkbox"/>

	Name	Full Name	E-Mail			
	hartler	Juergen Hartler	juergen.hartler@tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	testmaspectras	test maspectras	juergen.hartler@tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	hackl	Hubert Hackl	hubert.hackl@tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	stocker	Gernot Stocker	gernot.stocker@tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	prem	PremAnand Achuthan	premanand.achuthan@tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	montse	Montse Pinent	montserrat.pinent-armengol@TUGraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	guest	Guest Guest	mauri@sbox.tugraz.at		<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	rader	Robert Rader	robert.rader@tugraz.at		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Accept

Cancel

Transaction time = 1169 ms

Figure 11: Sharing in the AndroMDA environment. The data can be shared with users and institutes. When a user obtains a shared item he/she cannot share it with others. Read rights are granted by default whereas update and delete rights can be granted individually (last two columns of the table).

The database transaction of huge datasets via EJB was moreover slow. To address this issue, a generic module was developed to transfer a whole value object tree using direct JDBC statements within a single database connection. The performance gain is related to the amount of data. In some tests the time necessary for a complete database transaction was approximately one third of the time needed by EJBs.

The development infrastructure described above proved to be extremely useful. Development speed in particular was dramatically increased. The extensions developed for the AndroMDA template project are popular with our developers and are used in many of the applications developed at the Institute for Genomics and Bioinformatics (TAMEE: data management system for tissue microarrays [123]; SampleDB: data management system for tissue samples; PPIX: [124]; RTPCR: management system

for real-time polymerase chain reaction experiments (PCR) [125]; LicenseManager: management system for license keys; MammaDB: database for the clinical data of breast cancer patients; StocksDB: database for biological stocks [126]; mouseDB: laboratory mice information management system).

3.2 MA_{SS} SPECTRometry Analysis System (MASPECTRAS)

The aim of this thesis was to develop a uniform, scalable and extensible IT platform for the management and analysis of LC-MS/MS data that facilitates and accelerates proteomics research. This ambitious goal has been realized with the aid of previously described tools. The result, the MA_{SS} SPECTRometry Analysis System, (abbreviation: MASPECTRAS) encompasses a variety of tools, integrated in a single platform and based on a MIAPE compliant schema.

3.2.1 MIAPE compliance

The starting point for MASPECTRAS was the development of a database schema. At this time, proteomics data was captured in a standardized manner using the PEDRo schema [44]. MIAPE was still in a rudimentary form not ready for use. To begin with, PEDRo was remodeled as an UML diagram with entity beans compatible with AndroMDA.

The gradual evolution of the MIAPE specifications was subsequently tracked and new developments were incorporated into the application data model. The result is a fully MIAPE compliant schema, fulfilling the MIAPE standards for gel electrophoresis (GE v1.0), mass spectrometry informatics (MSI v0.7), mass spectrometry (MS v2.1), gel (image) informatics (GI v0.1), and column chromatography (CC v0.2) (for a description of the standards see 2.2; the MASPECTRAS database schema at Appendix C).

Report beans, Struts servlets and JSPs were gradually generated and manually adapted to highly sophisticated, flexible web-pages (see figures 12 and 13). The main goal was to design a series of user-friendly pages to facilitate data entry. The user can for instance link information between different database tables, without seeking the main menu for the corresponding input mask. Help features are furthermore provided to minimize misunderstandings. The system is designed to retrieve information rapidly whilst still providing the flexibility to store all types of proteomics experiments. The rapid development and incorporation of new standards depended critically on the AndroMDA code generation framework (see 3.1).

The screenshot displays the MASPECTRAS software interface. On the left is a 'Tree View' showing a hierarchical structure of sample preprocessing steps: Sample, Gel 1D, Band, Gel 2D, Spot, Lc Column, Fraction, MS Experiment, and another Fraction, Spot, Lc Column. The main area is titled 'New Gel2D' and contains a configuration window for a 2D gel electrophoresis experiment. The 'Buffer' tab is selected, showing settings for 'Buffer X' and 'Buffer Y'. The 'Running Buffers' section includes a dropdown menu set to 'testBuffer' and a 'Details for running buffer' dialog box. This dialog box contains a table with the following data:

Component	Concentration
comp1	1.0

Below this, there are sections for 'Add running buffer' and 'Add additional buffer', each with a dropdown menu and a blue button. The 'Electrophoresis Conditions' section has two conditions: Condition 1 (Voltage: 3 [kV], Voltage Mode: gradient, Time: 3 [h], Temperature: 37 [°C]) and Condition 2 (Voltage: 2 [kV], Voltage Mode: step and hold, Time: 2 [h], Temperature: 20 [°C]). At the bottom, there is a 'Buffer Equipments' section with a dropdown menu set to 'testEquipment' and a blue button. A 'Create' button is located at the bottom left of the configuration window.

Figure 12: MIAPE compliant input mask for 2 dimensional gels and navigation tree. The MASPECTRAS input for the sample preprocessing steps is kept completely flexible. All available steps can be combined. In this example (tree view), parts of the sample are preprocessed by a 1-dimensional gel electrophoresis experiment, other parts of the sample by a 2-dimensional gel electrophoresis experiment, and other parts by a liquid chromatography column, yielding bands, spots, or fractions respectively. One spot is preprocessed again by a liquid chromatography column delivering fractions, whereby with one fraction 2 mass spectrometry experiments are conducted. With the tree view the user can navigate quickly to the desired item. The right part of the picture shows parts of the input mask for 2-dimensional gel electrophoresis (the remaining input masks are at the remaining tabs). Although a lot of information is required, the input was minimized to as great an extent as possible. Whenever there is more information to enter (especially reusable information) then select fields are provided. To add items that are not present in the list, there is a link on the button, which leads directly to the corresponding input mask. When the information is entered the user returns to the input page where the blue button was clicked. The user is prevented from useless searching for the correct input page in the main menu. For detailed information about an item in a select field, the link on the button opens a closable box containing it. When several possible items can be entered, an “Add something” button (e.g. “Add running buffer”) is provided to add an additional input field. If it is convenient to have several input fields as for the “Electrophoresis conditions” they are displayed within the same page. With the button the additional fields can be removed again. The offers information about what must be entered at that field (normally the original MIAPE text appears).

Edit LcColumn

Edit Display Settings

Title: testLcColumn

LC Column Run Phases Run Settings Detection

Mobile phase components ?

Mobile phase component: mob comp1 1.0 ✕

Mobile phase component: mob comp2 2.0 ✕

Mobile phase component: mob comp3 3.0 ✕

Add mobile phase component

Gradient step 1: ?

Gradient Type: gradient 2.0 [min]

Purpose: preparative ?

Composition

Component: mob comp1 1.0 2.0 ✕

Component: mob comp2 3.0 4.0 ✕

Add component

Type: mob comp2 ?

Substance: testReagent ?

Time: 1.0 ?

Volume: 2.0 ✕

Add between run

Gradient step 2: ? ✕

Gradient Type: constant 2.0 [min]

Purpose: analytical ?

Composition

Component: mob comp2 3.0 ✕

Add component

Add between run

Add gradient step

Update

add Fraction

Nr.	FractionId	StartPoint	EndPoint	ProteinAssay
1	Fraction One	1.0	2.0	3.0 ✕

Edit Band

? Details for the starting point ✕

The x- and y-coordinates of the starting point.

Title: testRectangle

Area: 1.0 ?

Intensity: 2.0 ?

LocalBackground: 3.0 ?

Annotation: Annotation ?

AnnotationSource: AnnotationSource ?

Volume: 4.0 ?

Normalisation: ageNormalizationMethod ?

NormalisedVolume: 5.0 ?

LaneNumber: 6 ?

ApparentMass: 7.0 ?

Description: Description ?

LocalisationItem Type: Boundary Chain

Boundarypoints: ?

X: 1 Y: 2 ?

Directionstep: E 6 [pxs] ✕

Directionstep: NW 2 [pxs] ✕

Directionstep: SW 2 [pxs] ✕

Add direction step

Update

L 1 2 3 4 5 6 7 8 9 10 11 12

Figure 13: MIAPE compliant input masks for liquid column chromatography and for a band of a 1-dimensional gel electrophoresis experiment. The liquid chromatography input mask (left picture) demonstrates the versatility of the system; the mobile phase components entered at the top section appear in the drop down list of the select box. At the bottom a list of the fractions obtained from the liquid chromatography experiment is depicted. The picture of the band on a 1-dimensional gel electrophoresis experiment (right picture) demonstrates that the display of pictures is integrated, for when this facility is required.

3.2.2 Analysis pipeline

A standard compliant platform does not provide sufficient functionality to be accepted by the proteomics community. Features for the analysis of proteomics data must be provided. Figure 14 shows an overview of the tasks which can be carried out by MASPECTRAS. Researchers conduct their proteomics experiments as usual and generate mass spectrometry data. The data is searched by a database search engine of choice. Although from an informatics point of view the platform could start with a database search engine, this part has not been integrated for the following reasons:

- There is a lot of research going on in this area (see Appendix B) and yet another (additional) search algorithm was not desired.
- Many search algorithms are not free of charge and integration is problematic
- Integrating algorithms significantly increase maintenance effort, since they are frequently changed.

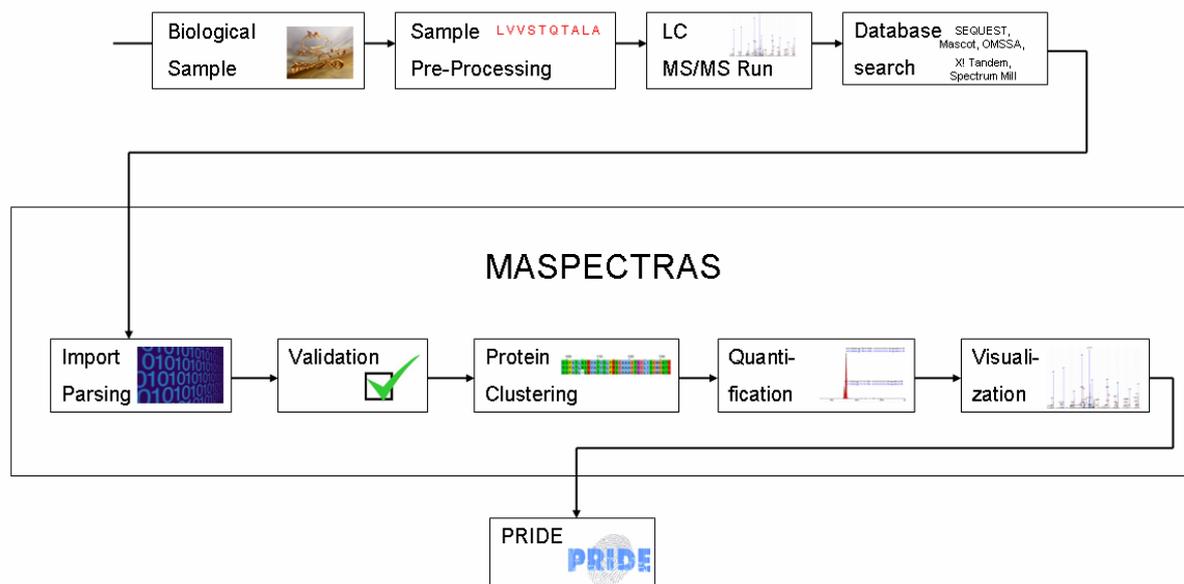


Figure 14: MASPECTRAS analysis pipeline. The mode of execution of mass spectrometry experiments is not influenced at all by the use of MASPECTRAS. Database searches are afterwards performed with the search engines of choice. The data is then imported to MASPECTRAS (see 3.2.3). There the search engine results undergo a further validation algorithm (see 3.2.4) to increase confidence in the results and to filter out false positives. The proteins are then clustered according to their similarities (see 3.2.5), and the remaining peptides are quantified (see 3.2.6). For the analysis of the resulting data, MASPECTRAS provides a variety of analysis tools. The final result can be exported directly from the view to a format for a public repository (PRIDE).

The analysis with the MASPECTRAS platform starts after the search engine results have been generated. MASPECTRAS provides several parsers, which support the native output formats of various search engines. The data is subsequently fed through an automatic processing pipeline and the results are scored by a well established algorithm (see 3.2.4). The identified proteins are then clustered according their sequence similarities (see 3.2.5), since the outcome of PFF experiments are peptides, and the proteins correspond just to a group sharing the same set of peptides. In a final step, the peptides are quantified by their chromatogram trace (see 3.2.6). The remainder of the analysis must be performed manually by the user, aided by different visualization tools provided by MASPECTRAS (see 3.2.7 and 3.2.8). The final results can be easily exported to different file formats with only a single mouse-click (see 3.2.9). One of the formats, the PRIDE-XML format, can be uploaded without further changes to the PRIDE data repository at EBI. The pipeline is implemented asynchronously using message driven beans (see 2.3) enabling the user to analyze uploaded data whilst the remainder is processed by the pipeline. The progress and the status of the import can be observed in an upload section.

3.2.3 Data import

The proteomics data necessary for the evaluation of a mass spectrometry experiment with MASPECTRAS comprises of the search engine results and the raw data from the mass spectrometry experiment. The raw data is needed for quantification, since the search engine result files store only the spectra for the found hits. In order to quantify a peptide, the chromatogram must be available: this cannot be reconstructed from the search engine results, only from the aforementioned raw data. The raw data is however stored by different mass spectrometers in proprietary formats (normally binary), which change continuously. Fortunately, converters relying on machine specific libraries, which generate files in commonly readable XML formats already exist. In recent years the two formats mzXML [127] and mzData [71] were established. Both of these are supported by the MASPECTRAS analysis platform. The native format of XCalibur 1.3 is furthermore also supported.

Every available search engine (see Appendix B) stores its results in a different proprietary file format. An attempt was made to establish a standardized storage format for search engine results (pepXML [46]). This format (or at least the converters) however lacks some important information (Second hits for instance are not stored for a spectrum. This is an issue since the first hit is frequently not the right one, especially in PTM research). This approach was therefore not acceptable. The aim was thus to establish a feasible platform, with search engine-independent implementation, which had to be generic and easily extensible to further formats. From our perspective the most commonly used search engines in the proteomics field were the commercial products SEQUEST, Mascot and SpectrumMill, and the freely available ones X!Tandem and OMSSA. Parsers for the native formats of these five were implemented, with the advantage that no file conversion is required.

A major goal of the thesis was to guarantee the comparability of results originating from different search engines, since these store their data in significantly different ways (apart from just using different scores). A standard presentation of the data was hence designed. One of the major challenges associated with this was the representation of the PTMs, since the way they are stored (sometimes with completely different names) varies greatly. Moreover it should be possible to query for modifications originating from different search engines (see 3.2.7). The only unifying feature is the difference in mass due to the PTM. To address this, a lookup table was introduced, containing all of the masses used for the database search, with identifiers for the modifications stored in the column MOD_NUMBER (see figure 15). As well as the sequence string a modification string is stored in the Peptidehit table. An unmodified amino acid is marked with 0 at the corresponding position of the molecule, a digit bigger than 0 is used if PTMs have taken place (the meaning of the value is stored in the lookup table).

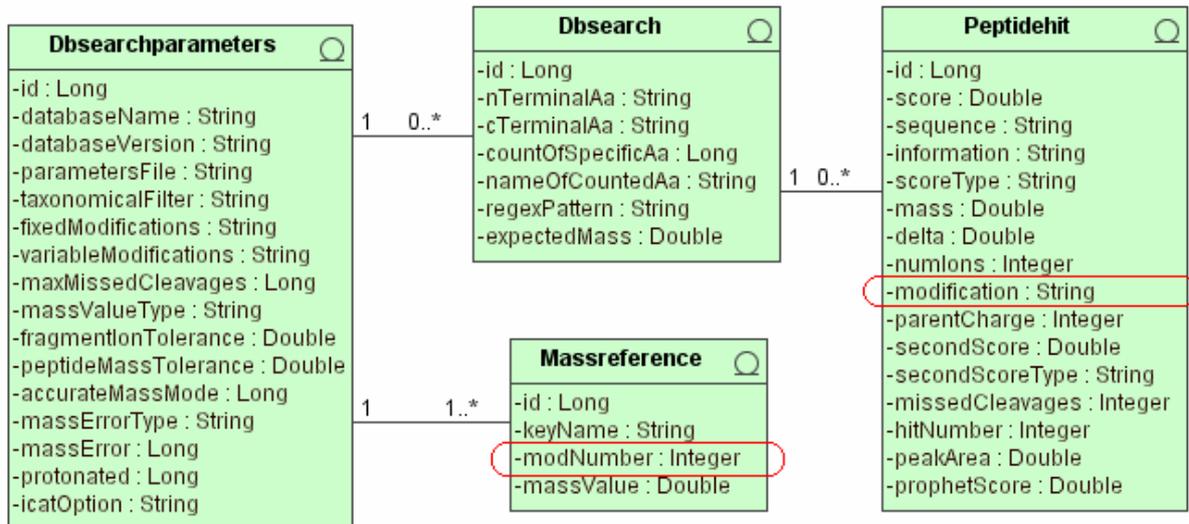


Figure 15: Schema for the storage of PTMs. As well as the sequence string a modification string is stored in the Peptidehit table. An unmodified amino acid is marked with 0 at the corresponding position of the molecule, a digit greater than 0 if PTMs have taken place. For the whole search, one table called Massreference serves as lookup. The lookup works via the field modNumber in the lookup table.

Another challenge was to compare proteins originating from different search engines since the protein sequences are not stored in the search engine result files. Only a non uniform identifier for the corresponding protein is stored. Even if the same protein sequence databases are used, the search engines extract different parts of the accession string (e.g. X!Tandem: gi|231300|pdb|8GPB; Mascot: gi|231300; SpectrumMill: 231300). Moreover different databases provide different identifiers (e.g. NCBI nr [128]: gi|6323680; MSDB [129]: S39004). To provide a universal mean of obtaining comparable identifiers from several database search engines using different protein sequence databases, a generic sequence database management module was implemented.

Since such a solution could be of interest for multiple applications, the management and parsing of sequence databases was implemented on the JClusterService side [116] with a corresponding API, while the web-interface was implemented directly in MASPECTRAS (see figure 16). For MASPECTRAS it was important to retrieve the protein sequence, as well as some additional information such as the description or the organism based on the identifier used by the search engine. The service provided by the JClusterService is designed to dynamically store parsing rules to extract almost every part of the header string with a corresponding regular expression [130].

Database yeast

Rule to parse accession string from Fasta file:	(g \d+ \
Rule to parse description string from Fasta file:	[^\]* (.*)
Rule to parse organism from Fasta file:	[^\]*([w s +])*

Nr.	Databasename	Version	Status		
1	yeast	04090683289	Active	✔	📄
2	yeast	04090683291	Active	✔	📄
3	yeast	04090683290	Inactive	✔	📄
4	yeast	1	Active	✔	📄

Return

=====

Complete Entry:

```
>gi|19114688|ref|NP_593776.1| hypothetical homeobox domain protein [Schizosaccharomyces pombe]Dgi|1723488|sp|O10328|YD73_SCHPO Hypothetical
MRSYSNPENGGQINDNINYSKRP TMLPENLSLSNYDMSFLGQFSDNNMQLPHSTYEQHLQGEQONP TNPYFPEFD
ENKVDWKQKPKPDAPSFADNNSFDVNSSSLTNPSPVQPNIVKSESEP ANSKQNEVVEATSVEKAKENV AHESGTPESG
GSTSAPKSKRQLTADQLAYLLREFSKD TNPPIAREKIGRELNIPERSVTIWFQNRRAKSKLISRRQEEERQRLREQR
ELDSLNRKVSQAFAHEVLSTSPYVGGIAANRQYANTLLPKPTRKTGNFYKSGPHQSSMEPCIAESDIP IRQSLST
YNSLSPNAVVSQRKYSASSYS AIPNAMSVSNQAFDVESPPSYATPLTGIRMPQESDLYSPREVSPSSGGYRMPF
HSKPSYKASGVPVPPMATGHRMRTSSEPTSYDSEFYFSC TLLVIGLWKLRASPDLMCFYSPPKLFA YLIQFQGIQ
YRIYSFFVIESIHVFRVEEPLINELSATASSRDKPAPNEYWLQMDIQLSVPPVPHMITSEGGNCTDFTEGNQASEVLL
HSLMGRATSFQMLDVRVRASPELGSVIRLQKGLNPHOFLDPQWANQLPROPDSVFDHQRNPP IQGLSHDTSSEYGNK
SQFKRLRSTSTPARQDLAQHLLPKMTNTEGLMHAQSVSP ITQAMKANVLEGSSTRLNSYEPSVSVSAYPHHNLALNDMT
QFGLGTNINISYPLASPDVGLPRASNSPSPRVMPHTQGINTE IKDMAAQFNSQTGGLTPNSUSMNTNVSVPVFTQON
REFGGIGSSSISTTMMAPSQQLSQVFPDVS LATENSVPYGFVPEVSESVYAQARTNSSVSAGVAPRLFIQTPSIPLAS
SAGQDNLIEKSSGGVYASQPGASGYLSDQSGPFDVYSPSAGIDFQKLRGQQFSPDMQ
```

Rule accession_rule: gi|19114688,gi|1723488,gi|7490714,gi|1213267,
Rule description_rule: hypothetical homeobox domain protein [Schizosaccharomyces pombe],Hypothetical protein C32A11.03c in chromosome I,hypc
Rule organism_rule: Schizosaccharomyces pombe,null,null,Schizosaccharomyces pombe,
=====

Complete Entry:

```
>gi|496693|emb|CAA56020.1| B-127 protein [Saccharomyces cerevisiae]
MFFSFLAQFFPKISSHSLGWNSPGRGSHGNLNVFWYKLSISGLIEEDIVVDSPGFVVISLILLVVEVDGLLIVLFFV
AFVPGFATVVP IPLKLENVFLGDIWVVDVGLDSSDVLSSIVFIPGL
```

Rule accession_rule: gi|496693
Rule description_rule: B-127 protein [Saccharomyces cerevisiae]
Rule organism_rule: Saccharomyces cerevisiae
=====

Complete Entry:

```
>gi|6323056|ref|NP_013128.1| AICAR transformylase/IMP cyclohydrolase; Ade16p [Saccharomyces cerevisiae]Dgi|1709914|sp|P54113|PU91_YEAST Bifu
MGKYTKTAILSVYDKTGLLDLAKGLVENNVRILASGGTANMVREAGFPVDVSSITHAPEMLGRRVKTLPVAVHAGILAR
```

Figure 16: Generic protein sequence database administration. At the top parsing rules for the accession, description and organism are defined via regular expressions. When the parsing rule definition is tested, the parser applies the rules to the first ten database entries and returns the results to check the parsing rules manually. The complete entry is shown together with parts of the entry extracted by the rules. If one entry has several accession strings and descriptions they are comma separated.

In order to compare results originating from different protein sequence databases, the database schema shown in figure 17 was developed. The presence of a given protein in a database can be quickly established via the lookup of the checksum. If the accession string differs, a new entry referring to the stored protein sequence is stored. When results from different searches need to be compared, the same proteins are detected quickly, since, although they can originate from different protein sequence databases, they share the same protein sequence identifier.

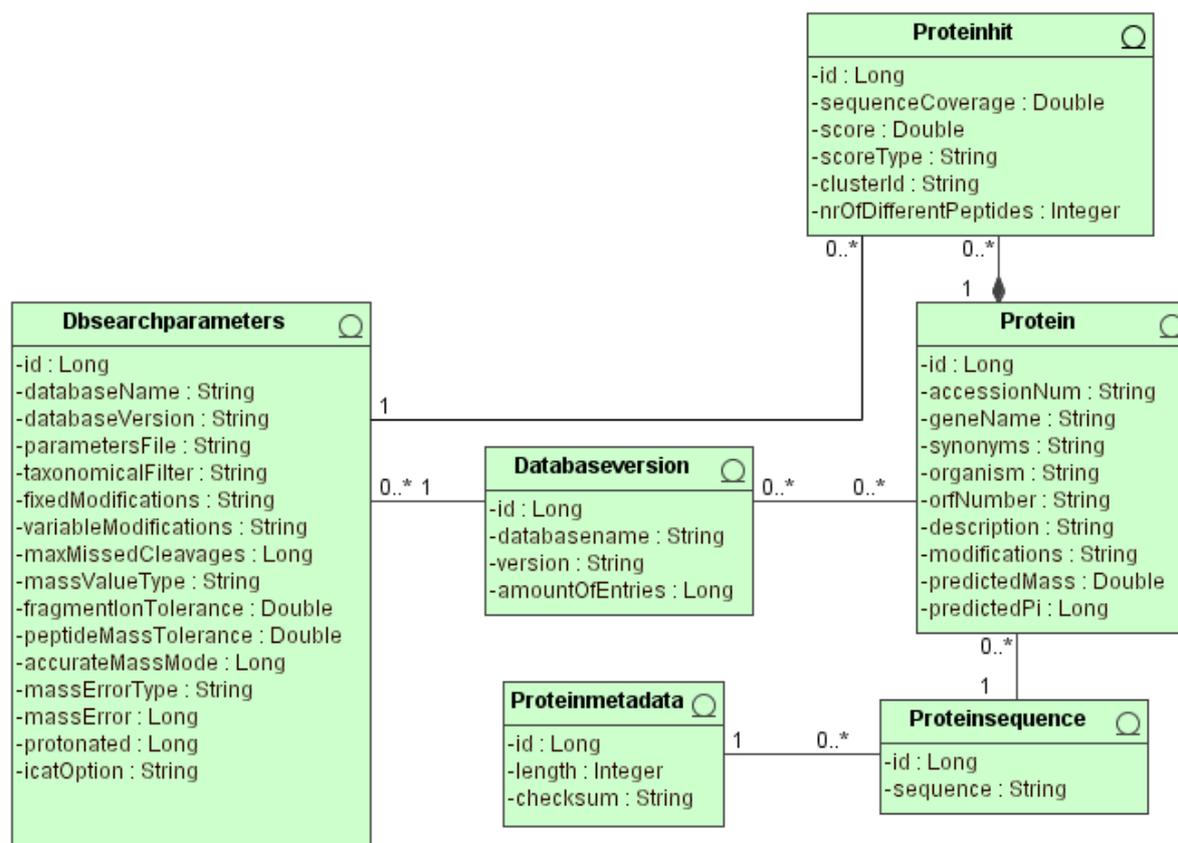


Figure 17: Schema for the storage of proteins. The table Proteinsequence contains the amino acid sequences. Proteinmetadata serves as lookup table when data is imported to MASPECTRAS. There the length and a MD5 checksum of the protein sequence are stored to find the same sequences quickly. The Protein table stores annotations of the specific database. Proteinhit stores parameters specific for a specified database search.

3.2.4 Data validation

Since the database search algorithms are based on statistical methods the detection of false positives is adherent. Several approaches for the re-scoring of peptide hits have been published [12,13,131], motivated by the desire for increased reliability and less manual effort. In a benchmarking study from Kapp et. al. [25] the specificity and sensitivity of PFF search algorithms was tested. The PeptideProphet [12], a validation algorithm based on SEQUEST or Mascot search results, turned out to be the most reliable algorithm. MASPECTRAS therefore provides an in-house translated Java-derivative of the original C++ algorithm. The statistical model of the PeptideProphet algorithm incorporates a linear discriminant score based on the database search scores (for SEQUEST: XCorr, dCn, Sp, rank and mass difference) as well as the tryptic termini and missed cleavages.

After the search results entered MASPECTRAS, the results from SEQUEST and Mascot are automatically validated by the PeptideProphet module. The data must in addition pass user-definable filters. The filters can be applied to the native search engine scores (in combination with the charge state, since the search engine scores are dependent on the charge state of the peptide) as well as to the newly calculated PeptideProphet scores for SEQUEST and Mascot respectively. This filter should

remove the most unlikely data, which is not to be stored in the database (reduction of quantity). The filter threshold should be kept rather low, since filtering at the visualization phase is still possible and the apparent false positive data is still accessible.

3.2.5 Protein clustering

Before analysis by mass spectrometry, proteins are broken down to peptides (normally by an enzymatic digestion). The immediate results of such experiments are thus just peptides, and not proteins. The peptides are mapped to proteins, which contain the amino acid sequence of the peptide. Biological protein sequence databases contain a lot of homologue proteins, and the fragments (peptides) of the proteins could be the same by chance. Thus mass spectrometry experiments do not identify proteins but instead identify a group of proteins, which, due to sequence homologies, share the same or a similar set of peptides.

Protein clustering is an essential feature of an integrated proteomics platform. A clustering approach has already been developed in-house for a large-scale comparative transcriptomic study (more precisely cDNA microarrays) [132]. The clustering algorithm is based on Markov Clustering [133] using BLAST [134]. The clustering pipeline has been integrated into the JClusterService (see 2.3) [116], because parallel computing proved to be quite effective for this purpose. Some slight modification of the existing service was required for the protein comparisons needed for MASPECTRAS.

After the proteins pass the filter described in 3.2.4, a file in FASTA format is assembled containing all sequences requiring clustering. Each sequence is then compared to all other sequences. The all-against-all sequence similarities generated by this analysis are parsed and stored in an upper triangular matrix. This matrix represents sequence similarities as a connection graph. Nodes of the graph represent proteins, and edges represent sequence similarity that connects such proteins. A weight is assigned to each edge by taking the average pair wise $-\log_{10}$ of the BLAST E-value. “These weights are transformed into probabilities associated with a transition from one protein to another within this graph. This matrix is passed through iterative rounds of matrix multiplication and inflation until there is little or no net change in the matrix” [133]. The final matrix reveals protein clusters. The identifier of the corresponding cluster is stored for every protein hit.

3.2.6 Peptide quantification

The steps of the analysis pipeline described above provide a qualitative idea, of which peptides or corresponding proteins are present in a biological sample. They however do not allow conclusions regarding the amount of proteins present in the sample to be drawn because the ions measured are put

on an exclusion list for a definable period of time (see Appendix A). The reason for this is that less abundant proteins would otherwise be lost. To gain quantitative information the raw data from the mass spectrometer must be analyzed. The virtual chromatograms are calculated from the raw data; these are then smoothed and afterwards used to calculate the peak area (see 2.1).

In order to integrate a mass quantification in MASPECTRAS existing solutions were reviewed [26,27]. The results published by ASAPRatio [26] looked promising and fitted largely to the expectations of how to implement such a module. In order to be able to implement improvements of the algorithm it was reprogrammed for the Java programming language. In our implementation the m/z range for the chromatogram is user-definable. The chromatogram of one charge state is calculated by the summation of the ion intensities, smoothed tenfold by repeated application of the Savitzky-Golay smooth filtering method [135]. For each isotopic peak, center and width are determined. The peak width is primarily calculated by using the standard ASAPRatio algorithm and for further peak evaluation an additional algorithm for recognizing peaks with saddle points has been implemented. With this algorithm, a valley (a local minimum of the smoothed signal) is recognized to be part of the peak and added to the area. The calculated peak area is determined as the average of the smoothed and the unsmoothed peak. Background noise, which is estimated from the average signal amplitude of the peak's neighborhood (50 chromatogram value pairs above and below the respective peak's borders), is subtracted from this value. The peak error is estimated as the difference between the smoothed and the unsmoothed peaks. A calculated peak area is accepted when the calculated peak area is bigger than the estimated error and the peak value is at least twice the estimated background noise. The peak area is otherwise set to zero. The calculation takes place automatically in the course of the analysis pipeline of MASPECTRAS. The peptides identified are combined into groups (peptides having the same sequence and same modification). These groups are then further subdivided according to their charge state. For each subgroup the median over the masses of the found peptides is calculated. For the calculation of the chromatogram this median is taken as the center of the m/z range, and not the *in silico* calculated ideal value. The reason for this approach is that the results generated by mass spectrometer are subject to variable error that is dependent on the instrument that is used. Normally, the error in m/z direction remains more or less constant for a given peptide. Despite this, we chose the median, because it allows more robust identification of outliers and false positives. The calculation can take place in MASPECTRAS directly or on a computing cluster (see 2.3), according to the number of peptides requiring quantitation. The threshold for job delegation can be set in a configuration file. A threshold is useful because the transfer of big MS raw data files is time-consuming and not feasible for a small number of peptides. Starting with approximately 50 peptides the gain in time increases almost in linear proportion to the number of processors used. After the calculation is finished, the retrieved peak areas are assigned to the peptides in the database and permanently stored. This module has been implemented as an adduct to the rest of the pipeline. The data can be analyzed by the user during the quantification process.

3.2.7 Visualization – views on the data

One of the goals for the development of this platform was to provide a uniform set of analysis tools compatible with the most commonly used search engines. Another key goal was that the searches from different search engines should be comparable or even better mergeable. These features endow MASPECTRAS with true novelty and render it superior to all other available mass spectrometry applications.

MASPECTRAS provides 3 merged views for the visualization of the data:

- The peptide view shows the peptides (see figure 18). There are two forms of the peptide view: one shows the peptides of one protein, the other all the peptides of the selected searches.
- The protein view lists the proteins found in one search (see figure 19).
- The clustered protein view lists the group of proteins found and displays one protein as representative for the whole group (see figure 19).

Merging means that the same proteins/peptides originating from different searches are displayed in one table row (see figure 20). The merging of searches works via JDBC statements of report beans, which had been adapted to the specific problem. The query, assembled by the generic JDBC generator (see 3.1), is first of all executed. The retrieved data is then post-processed (e.g. the same proteins originating from different searches are grouped; proteins are grouped according to their cluster identifier for the clustered protein view). Proteins with the same sequence are merged (see 3.2.3). The criterion for the same protein cluster is fulfilled if two proteins originating from different clusters (more precisely searches) are the same. All of the proteins of the two clusters are then merged into one cluster. The criterion for the same peptide is met if two peptides have the same sequence and the same modifications. A modification is considered to be the same when the mass shift is equal up to the second decimal place.

Most of the filters for the views were implemented through generic JDBC statements. The creation of all the required filters was however not possible. Then filtering is performed via post-processing in the report beans.

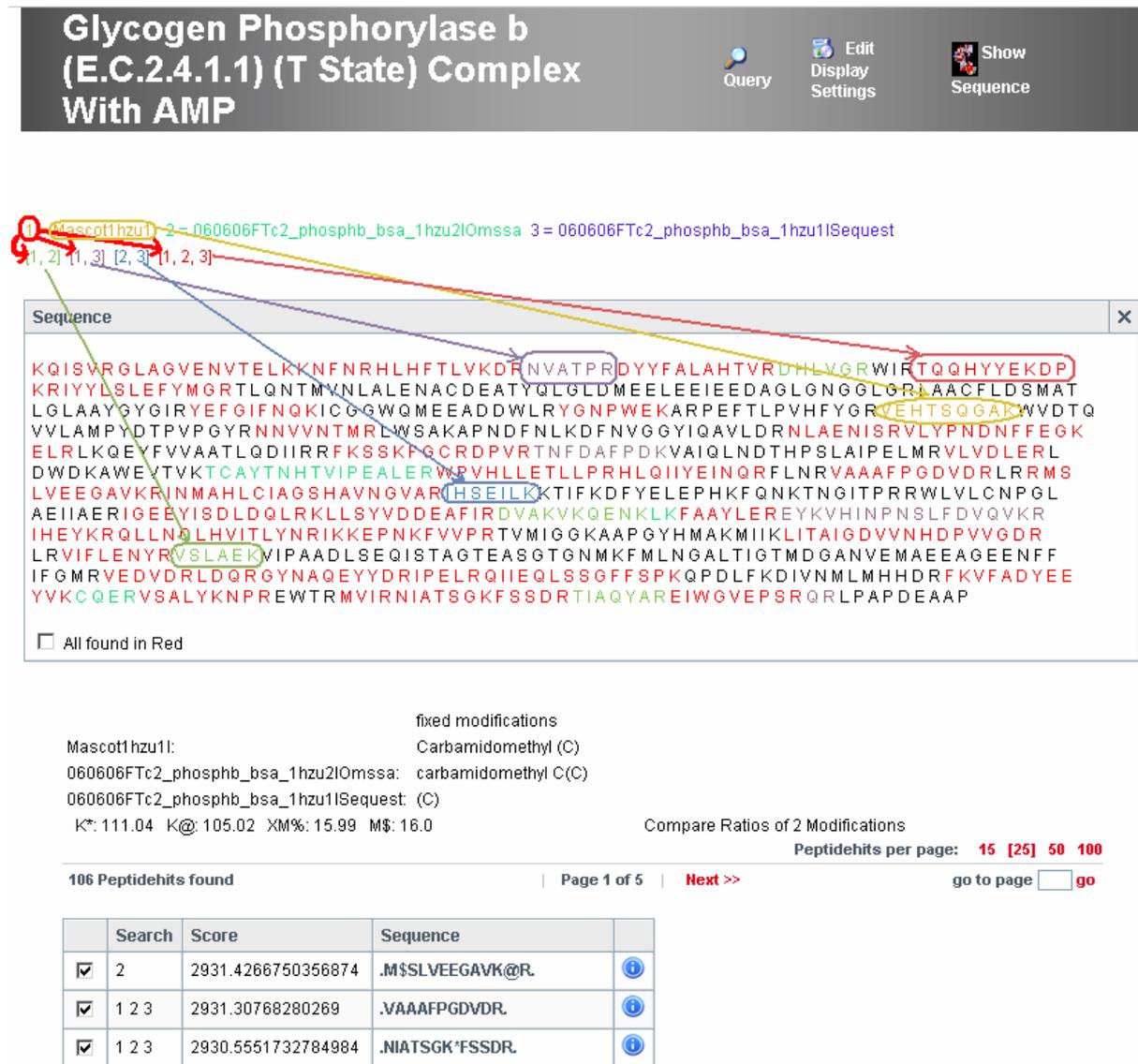


Figure 18: Peptide view. There are two versions of this. One of them lists all the peptides found, the other one only the peptides for a selected protein. Here the protein version is depicted. The description of the protein is displayed at the top. Below the header bar the searches are listed in one row; they receive an identifier for the list (1-3) and they are color encoded. Below, the possible combinations of identifiers are listed, and they are color encoded as well. The next box shows the sequence of the protein. The sections found by one or several search engines are color-encoded correspondingly. With the numbers of searches the number of combinations rises exponentially. The combinations are therefore only displayed for up to 4 searches. The sequences found by only one search engine have a separate color; the others are displayed in red. Underneath the sequence box the peptide modifications and their encodings are shown i.e. the searches have a fixed carbamidomethylation on cysteine and “K*: 111.04” for the variable modification means that on lysine (K) a modification of +111.04Da is possible. Whenever the modification occurs a “*” is placed behind K. The found peptides are listed below.

Protein View

1 = Mascot1hzu11 (Partitioning 
 2 = 060606FTc2_phosphb_bsa_1hzu2Iomssa (Partitioning 
 3 = 060606FTc2_phosphb_bsa_1hzu1ISequest (Partitioning 

Proteins per page: 15 [25] 50 100

8 Proteins found | Page 1 of 1 | go to page go

Nr.	Search	AccessionNum	Organism	GeneName	SequCovMax	Score	Amount of Peptides
1	1 2 3	gi 231300		Glycogen Phosphorylase b (E.C.2.4.1.1) (T State) Complex With AMP	53.97	43176.19	84
2	1 2 3	gi 162648	Bos taurus	albumin [Bos taurus]	39.71	29706.27	35
3	1 2 3	gi 418694	validated	serum albumin precursor [validated] - bovine	38.06	29623.04	34
4	1 2	gi 999627		Chain B, Porcine E-Trypsin (E.C.3.4.21.4)	21.96	108.41	2
5	1 2 3	gi 136429		Trypsin precursor	7.8	108.41	2
6	1 2	gi 3318722		Chain E, Leech-Derived Tryptase InhibitorTRYPSIN COMPLEX	8.08	108.41	2
7	1	gi 39794653	Homo sapiens	Keratin 1 [Homo sapiens]	2.18	81.04	1
8	1 3	gi 1346343		Keratin, type II cytoskeletal 1 (Cytokeratin 1) (K1) (CK 1) (67 kDa cyto keratin) (Hair alpha protein)	2.18	81.04	1

Proteins per page: 15 [25] 50 100

8 Proteins found | Page 1 of 1 | go to page go

Export Current View: Excel | DOC | TEXT | PRIDE XML

<< To Cluster View

To Peptide View >>

Clustered Protein View

1 = Mascot1hzu11 (Partitioning 
 2 = 060606FTc2_phosphb_bsa_1hzu2Iomssa (Partitioning 
 3 = 060606FTc2_phosphb_bsa_1hzu1ISequest (Partitioning 

Proteins per page: 15 [25] 50 100

4 Proteins found | Page 1 of 1 | go to page go

Nr.	Search	AccessionNum	Organism	GeneName	SequCovMax	Score	Nr. of Proteins	Amount of Peptides	
1	1 2 3	gi 231300		Glycogen Phosphorylase b (E.C.2.4.1.1) (T State) Complex With AMP	53.97	43176.19	1	84	
2	1 2 3	gi 162648	Bos taurus	albumin [Bos taurus]	39.71	29706.27	2	35	
3	1 2 3	gi 999627		Chain B, Porcine E-Trypsin (E.C.3.4.21.4)	21.96	108.41	3	2	
4	1 3	gi 39794653	Homo sapiens	Keratin 1 [Homo sapiens]	2.18	81.04	2	1	

Proteins per page: 15 [25] 50 100

4 Proteins found | Page 1 of 1 | go to page go

Export Current View: Excel | DOC | TEXT | PRIDE XML

Details from								X
Nr.	Search	AccessionNum	Organism	GeneName	SequCovMax	Score	Amount of Peptides	
1	1 2 3	gi 136429		Trypsin precursor	7.8	108.41999999999999	2	
2	1 2	gi 999627		Chain B, Porcine E-Trypsin (E.C.3.4.21.4)	21.96	108.41999999999999	2	
3	1 2	gi 3318722		Chain E, Leech-Derived Tryptase InhibitorTRYPSIN COMPLEX	8.08	108.41	2	

To Protein View >>

To Peptide View >>

Figure 19: Protein view and clustered protein view. The protein view is a simple listing of the proteins found. The clustered protein view exploits the results of the protein clustering (see 3.2.5). For each protein cluster one representative (the highest scoring one according to the used sorting criteria) is shown. In this example the amount of proteins in the group for “Chain B, Porcine E-Trypsin (E.C.3.4.21.4)” is announced as 3. With the blue information button the members of this cluster are shown (box at the end of the page).

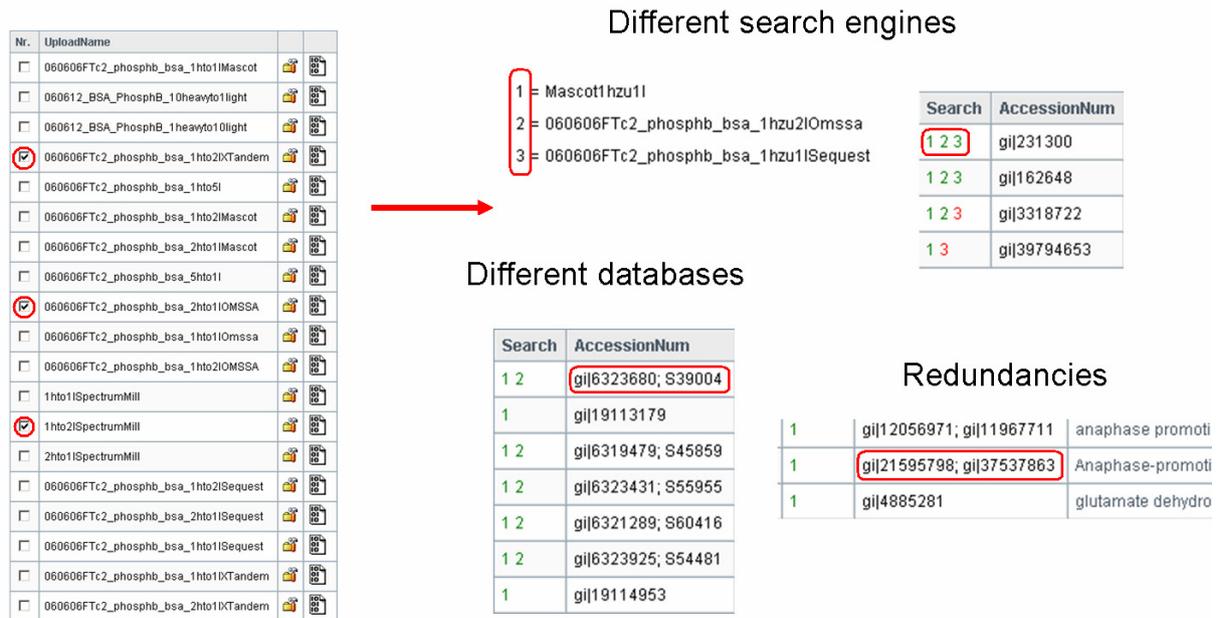


Figure 20: Merging of results in MASPECTRAS. MASPECTRAS can merge results from different search engines. In the column on the left side, a Mascot, X!Tandem, and SpectrumMill search is selected. The right side depicts what is merged. At the top the merging of different search engines is depicted. The searches are encoded with a number (1-3). If the corresponding protein has been found, the number is displayed for that search. In the center the merging of results from different databases is depicted. The two searches are conducted with Mascot, one with a NCBI nr database and the other one with MSDB database delivered with Mascot. MASPECTRAS recognized that the protein sequence is the same (see 3.2.3) and merged the entry, while both accession strings and description strings are displayed. A nice side-effect: the redundant sequences within the same database are recognized to be equal (“Redundancies” on the right side).

3.2.8 Visualization – additional visualization tools

Although the protein and peptide lists are the direct results of proteomics experiments, additional tools are necessary to raise confidence in the final findings.

Manual protein cluster validation

The Jalview Alignment Editor [136] has been integrated (see figure 21) for the visualization of the protein clusters (see 3.2.5) derived from a single search.

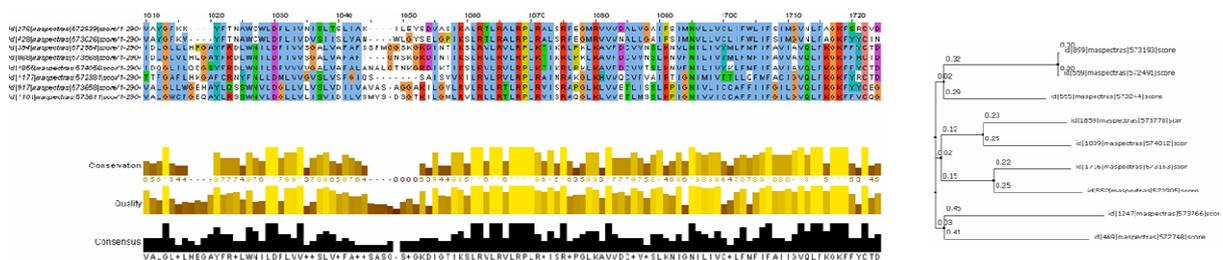


Figure 21: Jalview alignment editor. The protein sequences of one cluster are displayed in aligned form allowing easy detection of sequence similarities. A phylogenetic tree is furthermore provided to reveal the relationship between the proteins.

Spectrum viewer

Despite the availability of several computer-aided scoring algorithms, the manual visual inspection of an MS spectrum remains the most reliable indicator of the quality of an assignment. In view of this MASPECTRAS provides a spectrum viewer. This viewer provides some unique features which make it superior to other spectrum viewers (see figure 22):

- Ion series are selectable and unwanted ion series can be hidden. Custom settings can be stored for each user.
- Hit switching. The viewer can switch between the found peptide assignments for the spectrum.
- The spectrum view itself is a Java applet with zooming possibilities.
- The tooltips for the relative mass error help to identify unreliable series points.

Chromatogram viewer

Although the automatic quantification of a group of peptides can deliver reasonable results (see 3.4), the quantitation of a single peptide is error-prone. The reasons are: (i) additional peaks in a chromatogram in the m/z neighborhood; (ii) the peptides found are not in the main peak but in a smaller peak of the neighborhood. In view of these issues we have implemented a chromatogram viewer, which allows the manual inspection of the m/z neighborhood and manual correction of the automatic quantification (see figure 23). This feature is unique to MASPECTRAS.

Quantitative statistical section

A quantitative measurement for a single peptide or protein has no biological meaning. For this, the amount of the protein of interest must be determined in relation to another protein. Several labeling methods have evolved to compare proteins on the basis of mass spectrometry experiments. We therefore implemented a statistical module for the quantitative comparison of differentially labeled peptides. A ratio for each differentially labeled peptide pair of one protein is calculated. Based on these ratios, mean, standard deviation and a regression line is calculated (see figure 24).

The module provides some guidance about the quantitative relationship between the proteins, and has been of great value in the verification of the quantification algorithm (see 3.4). The quantification must be based on a comparison of the peptide ratios, since peptides differ in abundance due to their physiochemical properties.

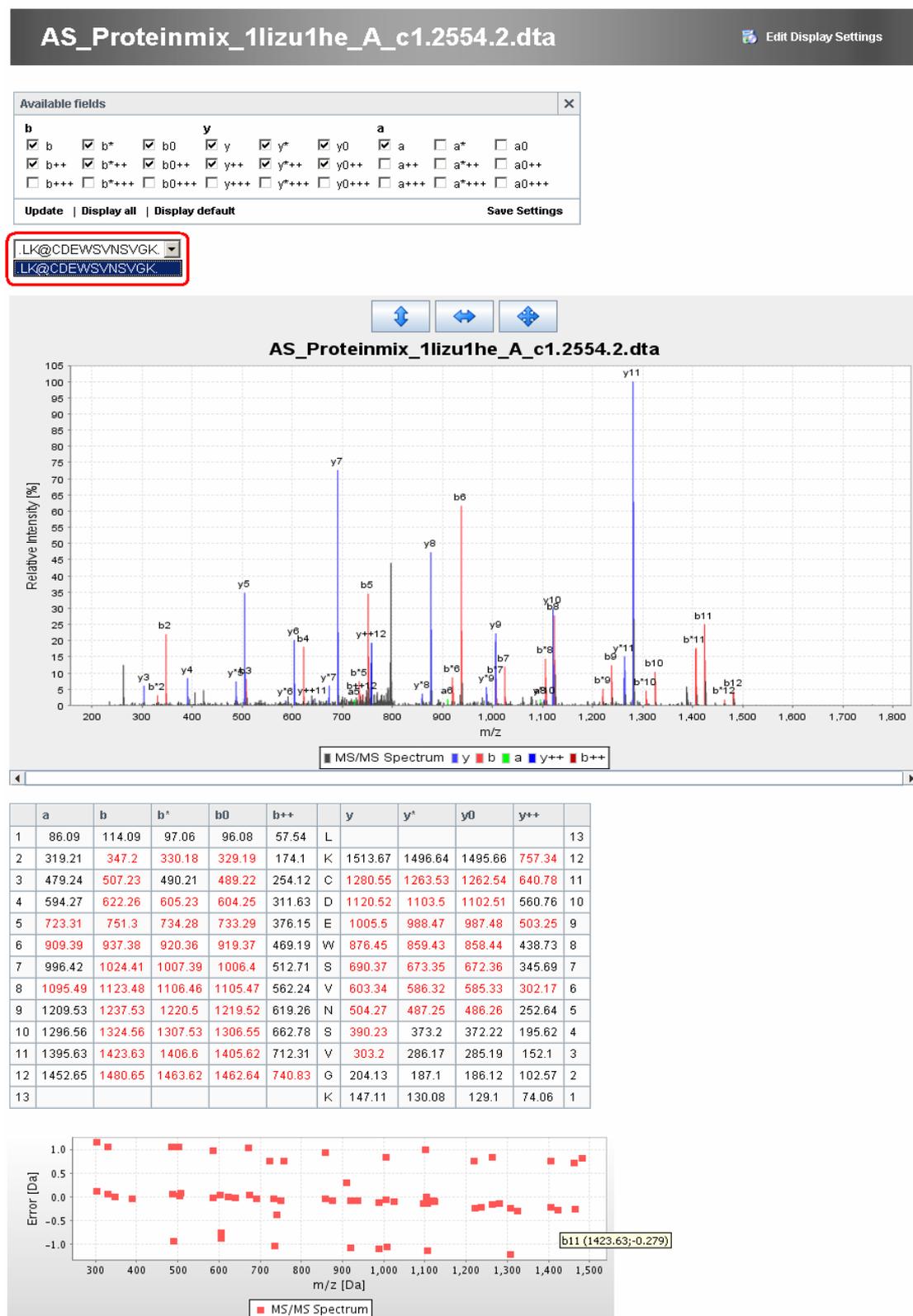


Figure 22: Spectrum viewer of MASPECTRAS. The ion series are selectable in the box with the check boxes (based on the standard nomenclature in proteomics [137,138]). With the select box it is easy to switch between first and second hit. The spectrum viewer is zoomable. In the box below the mass values for the corresponding ion series entries are calculated and the found hits are colored in red. The last box depicts the mass errors of the calculated ion series points. Every point has a tooltip to find the series entry.

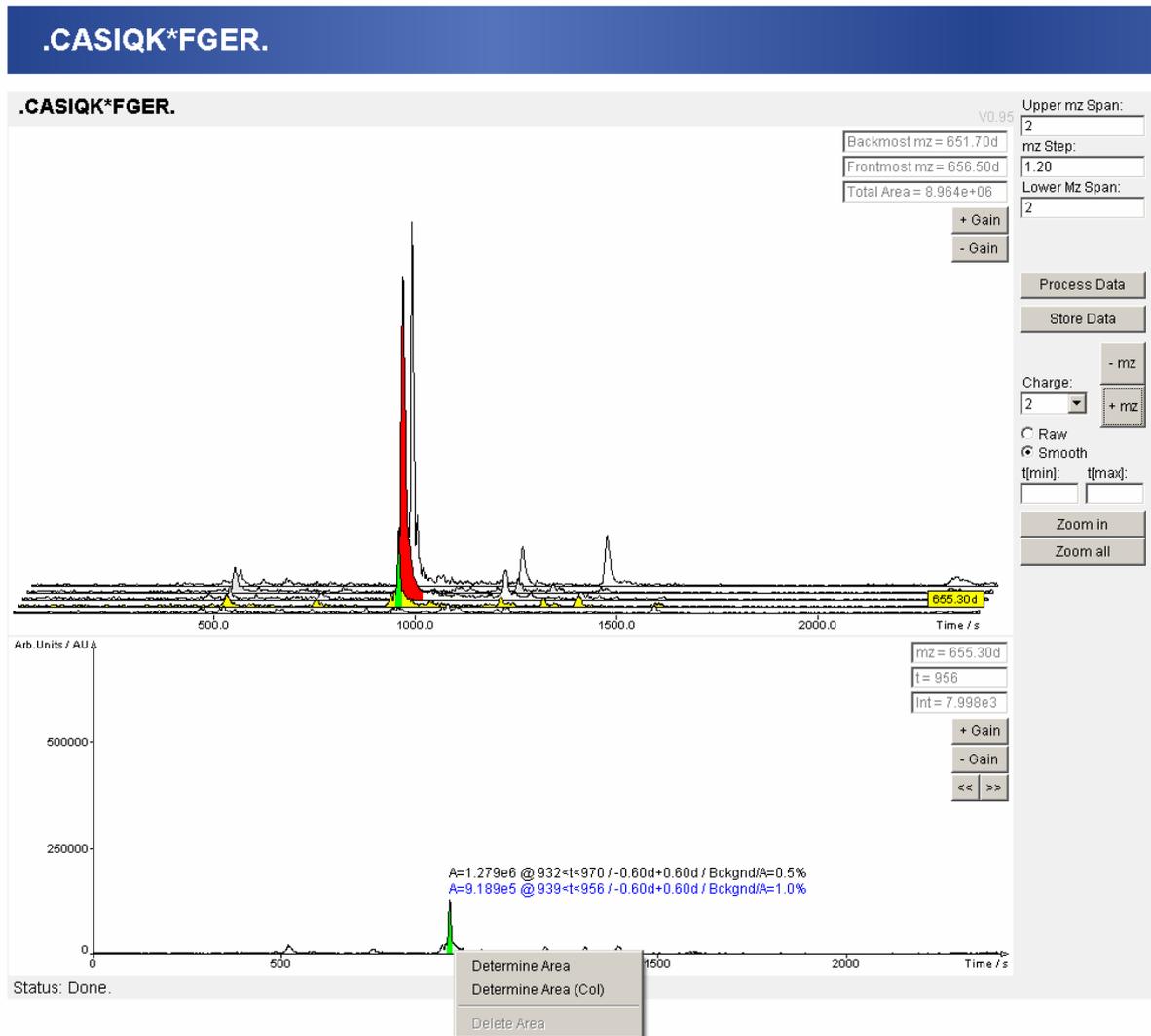


Figure 23: Chromatogram viewer of MASPECTRAS. Shows the chromatogram where the peak should lie plus some additional ones in the m/z neighborhood in a quasi 3 dimensional view. The selected chromatogram is displayed in the bottom 2-dimensional view. The m/z distance between the chromatograms is customizable (hence zooming in m/z direction is possible), as is the number of chromatograms displayed above and below. The viewer allows switching between the found charge states of a peptide. The red peak is one that is already stored in the database, while the green one is selected manually. Changes in the chromatogram viewer can be stored back to the database.

1=ICPL_Protmix_1lizu1he_A_c1_ms2
 2=ICPL_Protmix_1lizu1he_B_c1_ms2
 3=ICPL_Protmix_1lizu1he_C_c1_ms2

	Sequence	Z	N-termXK*: 111.04	N-termXK@: 105.02	Ratio 1/2	Ratio 2/1
<input checked="" type="checkbox"/>	.ALK*AWSVAR.	2	8856610.6171875	8761861.0	1.0108138690156692	0.9893018197047497
<input checked="" type="checkbox"/>	.ALK*AWSVAR.	2	9542612.75	9225358.75	1.0343893401435473	0.9667539689274303
<input checked="" type="checkbox"/>	.ALK*AWSVAR.	2	1.0170150546875E7	9550851.375	1.0648423001844691	0.9391061942475086
<input checked="" type="checkbox"/>	.CASIQK*FGER.	3	876386.25	958544.125	0.9142888961945284	1.0937461935305353
<input checked="" type="checkbox"/>	.CASIQK*FGER.	2	9323722.59375	9311852.25	1.0012747564535294	0.9987268664816393
<input type="checkbox"/>	.CLK*DGAGDVAQVK.	2	128293.22	661789.1	0.19385816417949464	5.15841055357407
<input type="checkbox"/>	.CLK*DGAGDVAQVK.	2	195198.17	634894.9	0.30744957945007906	3.2525658411654166
<input checked="" type="checkbox"/>	.YLGEYVK*AVGNLR.	3	506337.73046875	416672.78125	1.2151927201718316	0.8229147388725283
	Mean:				0.9666037807938672	1.0660752058942036
	Standard Dev.:				0.16904034	0.18780616

Export Current View: Excel | DOC | TEXT

fresh Areas

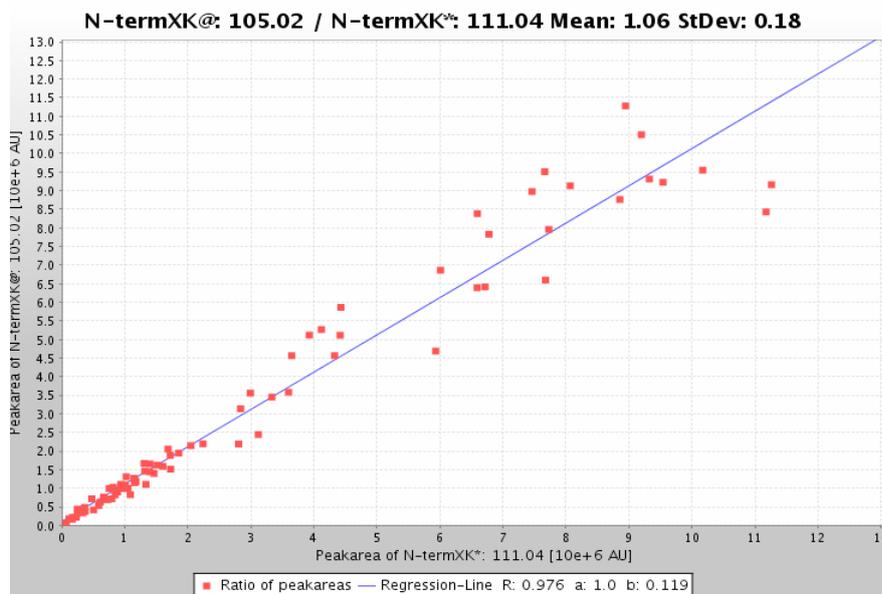


Figure 24: Quantitative comparison module. Compares the amount of peptides found with different modifications. The list shows the found peptides (of several searches) with the calculated peak areas of the different modification states, plus the ratio of the peak areas between each other. Additionally statistical parameters like mean, standard deviation, and a regression line are calculated. On the axes the peak areas are plotted. There the difference in abundance for single peptides is clearly visible (the peptides are originating from the same amount of protein).

3.2.9 Exporting

There are increasing demands to release the complete data sets generated by biological studies into the public domain. Since special emphasis was placed on standards compatibility in this project we utilized the associated advantages of easy export. A bar for the export (by one mouse-click) is provided in the protein and peptide view (see 3.2.7). As well as universal formats such as Excel, Word and text document, MASPECTRAS can export to PRIDE-XML also (see figure 25).

```
<?xml version="1.0" encoding="UTF-8" ?>
</spectrum>
<spectrum id="3192">
  <spectrumDesc>
    <spectrumInstrument mzRangeStart="120.0" mzRangeStop="1455.0" msLevel="2">
      <userParam name="basePeakMz" value="462.257" />
      <userParam name="basePeakIntensity" value="72.5922" />
      <userParam name="totalIonCurrent" value="1050.19" />
      <userParam name="precision" value="32" />
    </spectrumInstrument>
  </spectrumDesc>
  <precursorList count="1">
    <precursor msLevel="1" spectrumRef="3187">
      <ionSelection>
        <cvParam cvLabel="PSI" accession="PSI:1000232" name="Peak Intensity" value="1527.07" />
        <cvParam cvLabel="PSI" accession="PSI:1000040" name="Mass To Charge Ratio" value="480.15744" />
      </ionSelection>
      <activation />
    </precursor>
  </precursorList>
  <spectrumDesc>
    <arrayBinary>
      <data precision="32" endian="little"
        length="244">
        <data precision="32" endian="little"
          length="244">
          VT8ZLZant4yQNSH5j/ff7o/118P43eAEGkG0/c1CFQNuut96Y11Aj1D5P1HSB290X6U/CoEP2R1Y9mFhAMR1QV8u2J8mEvv/qcB3QJ4ndB9W1bM/pTEVP21B8Ua9od8G6T6H3vINT8s9GA/4HeWOK1v58Cg
          qM/ThsP8pp+z/CS2E/ouTeP88+46G3g98tkySP95Iz07HJJA1KcvLpE8Fz8ck/13E5Pw32j811vY/B3M/P61yveCoffA/B82CP9c8gT9Ffo0/YF1fAERH86+8z2fMYMPD/wfzgzLyF7M9Pndh18nZ6FAd
          hWpY3EJUADDDJqCVuXPMEK8EjEE/BP3KOP31t9mWkA21ghP3mHMD1vY/8g1dP97AAT+uo54PmHP/8Kd8hg2JAB8MPQ2L4J7Y8u+P7gB0H9FJ8z/c4/KFqP2rPUB8ms/NZE2P7M08Ke3hAc/qGQP
          mWd+KpNAMUM0PvYwD9e1CK/17c2Poc3KJutY1B23NYKcdj85zNAYG0bcvBKAqj18BQj1b0VKeekAn29E/0M62Pj+Kxj/DPF/FP70D7JkD+wR20/2LKP88tknG0PAZ825P8yeb804G1LARTxP6+JUH
          1MU/vv37Qioz3j8q9mAszG0DLNjB+448/sBcNEknk8ZyE1/Eb+W0G1nE8JULeAHH8QI2wz+VADzALNCp88/AEH8vU1AN5KHOPx1EC+Y1NAJ/vMPnUCUF7E7FBj+X00T117/vxNA5u8P86G1z+8BJM/
          2v0M8K1Y/C854/E0Y/P+IKQELwL5FC0y1F0vaxEH9c7AKCzE0P1786gJwA16gVQK970CP8H/P8HtQP1/K8AG6SE/PL0SOK6R2j8m/w8Ftc90F1B80A1U1p880U0kQskCKd0xADDKSP+q7z17uMys/Epu4P
          +98D8mK8AbhyH8P/uj8E8Ay8M8P8T76061mLm/u/vW0H3z8UCAM86zP28jgT/qtThAqz9P45C095F0M/eaRyG1t8H81BR8/7U8AP+2KbJ+8E8A8s1Vpx+51D+8K8g/87H8P66C1C1TSK/UmnVp+==</
          data>
        </data>
      </arrayBinary>
    </spectrumDesc>
  </spectrum>
  </mzData>
  <getFreeIdentification>
    <accession>I18694/Accession</accession>
    <database>karl/Database</database>
    <databaseVersion>04890683369/DatabaseVersion</databaseVersion>
  </getFreeIdentification>
  <peptideItem>
    <sequence>CCTKPESEK</sequence>
    <start>460</start>
    <end>468</end>
    <spectrumReference>338</spectrumReference>
    <modificationItem>
      <modification>4</modification>
      <modAccession>MOD:00625</modAccession>
      <modDatabase>PSI-MOD</modDatabase>
    </modificationItem>
    <additional>
      <userParam name="Observed" value="639.269433" />
      <userParam name="expt.Mass" value="1276.5310410000002" />
      <userParam name="expt.Mass" value="1276.5310410000002" />
      <userParam name="calc.Mass" value="1276.53484" />
      <userParam name="Delta" value="0.00402600000067" />
      <userParam name="Score" value="2.6985" />
      <userParam name="Rank" value="1" />
    </additional>
  </peptideItem>
  <peptideItem>
    <sequence>RHPYAVSVLLR</sequence>
  </peptideItem>
</mzData>
```

Figure 25: Excerpt of a PRIDE-XML document generated by MASPECTRAS.

For PRIDE-XML, not only the corresponding viewable list but all additional information associated with these results (spectra, the search parameters, information about the mass spectrometer etc) is included. The resulting file can be directly uploaded to the PRIDE proteomics data repository without any additional effort.

3.3 Validation study using large-scale data

In order to demonstrate that MASPECTRAS is a valuable platform for large-scale proteomics analysis, the system has been validated using data described in the study of Kislinger [139]. Their investigation focused on different organelles of different organs. The data from the heart cytosol compartment comprised of 84 Sequest searches and a total amount of 1.1GB. The database searches were performed against a protein sequence database containing equal numbers of decoy and normal proteins to test the reliability of their algorithm. Decoy proteins are the original proteins in an inverted amino acid orientation. The files were imported, parsed, the data analyzed (see table 1) and exported into PRIDE format. MASPECTRAS handled the large amount of data without any problems and the application of filter criteria similar to those used in the Kislinger study, took approximately 1.5 minutes. Differences in the number of proteins were caused by the use of different validation algorithms (MASPECTRAS uses PeptideProphet [12] while STATQUEST [131] was used in the Kislinger study). MASPECTRAS did not find any decoy proteins, while in the Kislinger study 0.3% have been reported.

Filter criteria	proteins found in both studies	proteins only by MASPECTRAS	proteins only by Kislinger	% found of Kislinger
2 Spectra for one Peptide + Prophet 0.95 + 1 First Hit	548	0	115	82.65%
2 Spectra for one Protein + Prophet 0.90 +1 First Hit	596	7	102	85.39%
2 Spectra for one Protein + Prophet 0.95 +1 First Hit	576	6	118	83.00%
2 first hit spectra for one Protein + Prophet 0.95	570	0	128	81.66%

Table 1: Comparison of the results of a large-scale study performed with MASPECTRAS with the published results [139]. The criteria chosen were as similar as possible, but the difference in the validation algorithm used caused different results.

3.4 Validation study using quantitative data

To validate the quantitative power of MASPECTRAS the mass spectrometry experiment described in 2.4 was conducted. Triplicates of differentially ICPL-labeled probes were mixed at 7 different ratios (1:1, 2:1, 5:1 10:1, 1:2, 1:5 and 1:10). To demonstrate the capabilities of MASPECTRAS, the quantitative analysis was performed with MSQuant [29], PepQuan (provided with the Bioworks browser from SEQUEST), and ASAPRatio implemented in MASPECTRAS. MSQuant is a program for the automatic quantitation of a set of peptides, while PepQuan provides a 2 dimensional chromatogram viewer. For the comparison to MSQuant the results of the automatic quantification with

MASPECTRAS were used. A manual correction with the integrated chromatogram viewer (see 3.2.8) was performed for the comparison to PepQuan. The performance of the quantification with ASAPRatio integrated in MASPECTRAS was superior to both other methods. Furthermore, for all ratios the relative error calculated was considerably lower than the relative error obtained with MSQuant and PepQuan (see table 2).

10 heavy to 1 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	22	33	27	40		
mean	4.64	8.94	8.31	9.85		
stdev	4.83	4.9	2.85	2.99		
CV %	104.09%	54.81%	34.30%	30.36%		

ratio: heavy/light

1 heavy to 10 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	20	82	28	39		
mean	7.54	7	9.77	9.29		
stdev	4.94	2.51	3.96	1.92		
CV %	65.52%	35.86%	40.53%	20.67%		

ratio: light/heavy

5 heavy to 1 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	14	43	50	53		
mean	2.94	4.27	4.16	4.67		
stdev	2.3	1.69	1.56	1.12		
CV %	78.23%	39.58%	37.50%	23.98%		

ratio: heavy/light

1 heavy to 5 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	16	67	41	40		
mean	13.36	3.74	4.25	4.84		
stdev	5.18	1.36	1.15	0.93		
CV %	38.77%	36.36%	27.06%	19.21%		

ratio: light/heavy

2 heavy to 1 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	25	50	48	72		
mean	1.048	2.17	2.07	2.03		
stdev	1.15	0.7	0.71	0.54		
CV %	109.73%	32.26%	34.30%	26.60%		

ratio: heavy/light

1 heavy to 2 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	16	74	42	47		
mean	4.24	2.07	2.11	1.94		
stdev	4.97	3.04	0.63	0.3		
CV %	117.22%	146.86%	29.86%	15.46%		

ratio: light/heavy

1 heavy to 1 light

	MSQuant		PepQuan		MASPECTRAS	
	auto	manual	auto	manual	auto	manual
# peptides	15	67	98	77		
mean	0.92	1.28	0.97	0.99		
stdev	0.46	0.48	0.24	0.19		
CV %	49.30%	37.50%	24.74%	19.10%		

ratio: heavy/light

Table 2: Quantitative evaluation of ICPL-labeled probes by MASPECTRAS, MSQuant and PepQuan. Automatic quantification by MASPECTRAS always performed better than with MSQuant. Similarly manual quantification by MASPECTRAS always performed better than with PepQuan. Automatic quantification with MASPECTRAS was moreover on most occasions better than manual quantification with PepQuan.

Chapter 4

Discussion

In this thesis an integrated and versatile bioinformatics platform for the management and analysis of proteomics mass spectrometry data has been developed: the MAss SPECTRometry Analysis System (MASPECTRAS). It is a web-based application accessible by common web-browsers with a relational database as storage back-end. The uniqueness of the platform lies in the MIAPE compliance, PRIDE export, and the scalability of the system for computationally intensive tasks, in combination with common features for data import from common search engines, integration of peptide validation, protein grouping and quantification tools (see table 3).

MIAPE compliance and PRIDE export are necessary to disseminate data and effectively analyze a proteomics experiment. As more and more researchers are adopting the standards, public repositories will not only enhance data sharing but will also enable data mining within and across experiments. Surprisingly, although standards for data representation have been widely accepted, the necessary software tools are still missing. This can be partly explained by the volume and complexity of the generated data and by the heterogeneity of the technologies used. MASPECTRAS is the first system which provides a completely MIAPE compliant database schema, capturing data from the design and at each step of the MS experiment to their evaluation and result export. It covers experimental design – sample generation, sample pre-processing, information about the mass spectrometry machine, and the analysis of the generated data. Nevertheless further changes to the standard are likely to occur and these should be tracked and incorporated. Furthermore, an automation of sample pre-processing steps would be feasible. Parsers for commonly used description files should therefore be implemented in future.

MASPECTRAS is moreover a system that provides the basis for consensus scoring between MS/MS search algorithms. It was recently suggested that the interpretation of the results from proteomics studies should be based on the analysis of the data using several search engines [25]. Importing and parsing the results from search engines and side-by-side graphical representation of the results is a prerequisite for this type of analysis. The system stores data retrieved from the mass spectrometry

machine and in addition the protein and peptide assignments from the widest spread tandem mass spectrometry (MS/MS) search engines (SEQUENT, Mascot, SpectrumMill, X!Tandem, and OMSSA).

	CHOMPER [15]	TPP [46]	GPM [47]	VEMS [48,49]	CPAS [50]	ProDB [51]	PROTEIOS [52]	GAPP [53]	PeptideAtlas [54]	EPIR [55]	STEM [56]	TOPP [57]	MASPECTRAS
Compliance													
MIAPE MSI compliant	—	—	*	—	—	—	✓	—	—	—	—	—	✓
MIAPE MS compliant	—	—	—	—	—	—	—	—	—	—	—	—	✓
MIAPE GE compliant	—	—	—	—	—	—	✓	—	—	—	—	—	✓
MIAPE GI compliant	—	—	—	—	—	—	✓	—	—	—	—	—	✓
MIAPE LC compliant	—	—	—	—	—	—	✓	—	—	—	—	—	✓
PRIDE export	—	—	—	—	—	—	—	plan.	—	—	—	—	✓
Data Import													
mzXML	—	✓	✓	conv.	✓	—	✓	—	✓	—	—	✓	✓
mzData	—	—	—	conv.	—	—	✓	✓	—	—	—	✓	✓
SEQUENT	✓	pepX	—	—	pepX	✓	—	—	✓	—	—	✓	✓
Mascot	—	pepX	—	✓	pepX	✓	✓	✓	—	✓	✓	✓	✓
SpectrumMill	—	—	—	—	—	—	—	—	—	—	—	—	✓
X!Tandem	—	—	✓	✓	✓	—	✓	✓	—	—	—	—	✓
OMSSA	—	—	—	—	—	—	—	—	—	—	—	—	✓
Data validation and visualization													
Search engine included	—	—	✓	✓	✓	?	—	✓	—	—	—	—	—
Additional validation algorithms	✓	✓	✓	✓	—	—	✓	✓	✓	✓	✓	✓	✓
Protein grouping - clustering	—	—	✓	✓	—	—	—	—	—	✓	—	—	✓
Merging of results from different search engines	—	—	—	?	✓	✓	✓	✓	—	—	—	—	✓
Customizable filtering	—	?	✓	—	part.	✓	?	SQL	—	✓	✓	?	✓
Spectrum viewer	✓	✓	✓	?	✓	?	?	—	—	—	✓	?	✓
Quantification													
Relative peptide quantification	—	✓	—	✓	—	—	—	plan.	—	✓	✓	✓	✓
Adjustable m/z width for quantification	—	—	—	—	—	—	—	—	—	—	—	—	✓
Visualization of chromatograms	—	✓	—	✓	—	—	—	—	—	?	—	✓	✓
Visualization of surrounding chromatograms	—	—	—	?	—	—	—	—	—	—	—	✓	✓
Scalability													
Parallel computing	—	—	—	—	—	—	—	?	—	—	—	—	✓

Table 3: Comparison of MASPECTRAS to other proteomics tools. ✓ fulfills criteria; — does not fulfill criteria; ? not enough information to answer this question; part. partially fulfills criteria; plan. planned; pepX fulfills via pepXML; conv. after conversion; * fulfills parts of MIAPE called XIAPE; MSI Mass Spectrometry Informatics; MS Mass Spectrometry; GE Gel Electrophoresis; GI Gel Informatics; LC Liquid Chromatography; SQL filtering over SQL only;

The search engine results data could furthermore originate from any protein sequence database in FASTA format since we integrated a generic sequence database management module, working with regular expressions. Since there is already an effective database management module, freely available search engines (X!Tandem, OMSSA) could in future be incorporated into the system to increase the integrity of the platform.

When the experimental data enters the system it must be processed by the MASPECTRAS analysis pipeline (see 3.2.2). This pipeline processes the data fully automatically to facilitate the rapid evaluation of proteomics data. In a first step the data from specific search engines is parsed into an internal search neutral representation which is MIAPE compliant. The data is afterwards revalidated by a derivative of the PeptideProphet algorithm [12]. PeptideProphet has been chosen because it performed well in a benchmarking study comparing different mass spectrometry algorithms [25]. In the next step, the proteins found are clustered according to their sequence similarities. To model these groups, an in-house developed clustering pipeline has been used. This tool uses the JClusterService [116], also developed in-house, to process the computing tasks in a parallel manner, and has already proven its worth in a comparative transcriptomic study [132,140]. The last step of the automated part of the MASPECTRAS proteomics pipeline is the quantification of the peptides found. In order to integrate a mass quantification in MASPECTRAS existing solutions were [26,27]. The results published by ASAPRatio [26] looked promising and fitted largely to the expectations of how to implement such a module. The ASAPRatio algorithm has therefore been integrated for quantification purposes. The quantification can be performed on the computing cluster as well. Tests showed that the computing time decreased linearly in direct proportion to the number of used processors. MASPECTRAS furthermore provides additional unique tools which include merged views (see 3.2.7), the chromatogram viewer (see 3.2.8) and PRIDE export (see 3.2.9). The integration of peptide validation, protein grouping, and quantification algorithms in conjunction with visualization tools is important for the usability and acceptability of the system.

MASPECTRAS is a fully web-based, scalable platform intended for deployment on high-performance server-machines, dedicated to the management of the growing amount of proteomics data. This approach provides the advantage that the client machines accessing the platform need less computing power and memory because the computationally intensive tasks are performed on the server. Furthermore the platform allows parallelization of time-consuming tasks (e.g. protein clustering and peptide quantification) since it features the in-house developed JClusterService [116], which is absent from all other existing proteomics systems (see table 3). The tasks can be spread on a customizable number of computing nodes to remove the computationally intensive tasks from the main application. MASPECTRAS is a 3-tiered platform that complies with the J2EE standard using EJBs and Jakarta Struts. The main advantages of the 3-tier approach are: (i) it is easier to maintain due to the separation;

the components can be more easily modified or exchanged; (ii) the load can be balanced better when the business logic is separated from the database. Another architectural milestone of MASPECTRAS has been created using the experience gained during the development of previous J2EE applications: a sophisticated AndroMDA code generator, meeting the requirements of an efficient UML-based development environment. The newly implemented AndroMDA extensions provide everything needed to generate highly sophisticated Struts web-pages, providing input pages and list pages with the amenities of querying, sorting, turning pages, and a customization of the display. Furthermore this code generation framework is directly linked to the in-house developed user management system (AAS). The AAS provides a clean concept to implement a multi-user environment, whereby security issues are handled at a centralized location. The combination of chosen technologies has proven to be successful for the development of this proteomics platform. The UML-based approach in particular simplifies the maintenance of the system. Since MIAPE is still evolving on an ongoing basis, it is of that utmost importance that the system is highly flexible to enable the implementation of adaptations with the minimum of programming effort.

Tests with datasets from a biological study showed that MASPECTRAS is well suited to analyze large-scale studies (see 3.3). The application of corresponding filter criteria worked in a fast and efficient manner. The discrepancies in results caused by the application of different validation algorithms nevertheless showed that application functionality could be extended by additional algorithms. Furthermore, the development of a consensus scoring algorithm, or probably a new scoring algorithm based on the quality of the spectra stored in the MASPECTRAS database should be considered. Despite these issues however, the quantitative verification of the software, indicated that the performance of MASPECTRAS was superior to other applications. Throughout the analysis some misquantifications by the automatic quantification were detected (see 3.2.8) mainly as a result of inaccurate peak detection. The main problem lies in the fact that a chromatogram is 2-dimensional, sometimes the third dimension (m/z value) is not detected properly. 3-dimensional peak detection should resolve this problem. The statistical section for the quantitative evaluation should furthermore be enhanced. An overview should be provided to detect up- and down-regulated protein ratios immediately. Moreover, the current peak quantification relies on the peptides found by database search engines. The probability that a given protein is not selected for MS/MS by the mass spectrometer or that is missed by the search engine algorithm is highest for biologically down-regulated proteins. Therefore, if one of the labeled proteins was detected, the differentially labeled counterpart should be quantified to compare the ratios. This could however lead to a higher rate of false positives, and further investigation in this direction will thus be necessary.

In summary, this is the first time that a comprehensive platform has been developed for the management of proteomics data in a MIAPE compliant manner. MASPECTRAS (i) provides the

amenities needed for analysis, (ii) features an automated analysis pipeline and unique analysis tools, (iii) provides an easy export functionality for the submission of the data to public repositories and (iv) is capable of managing the growing amount of mass spectrometry data in a scalable manner using parallel computing. The platform can be used for large-scale approaches and should become an indispensable tool for the rapid analysis of mass spectrometry datasets.

BIBLIOGRAPHY

1. Pandey A, Mann M: **Proteomics to study genes and genomes.** *Nature* 2000, **405**:837-846.
2. Graves PR, Haystead TA: **Molecular biologist's guide to proteomics.** *Microbiol Mol Biol Rev* 2002, **66**:39-63.
3. Hunter T: **Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling.** *Cell* 1995, **80**:225-236.
4. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422**:198-207.
5. Hunt DF, Michel H, Dickinson TA, Shabanowitz J, Cox AL, Sakaguchi K, Appella E, Grey HM, Sette A: **Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad.** *Science* 1992, **256**:1817-1820.
6. Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J, Bluggel M: **Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data.** *Proteomics* 2004, **4**:619-628.
7. Mann M, Wilm M: **Error-tolerant identification of peptides in sequence databases by peptide sequence tags.** *Anal Chem* 1994, **66**:4390-4399.
8. Lin D, Tabb DL, Yates JR: **Large-scale protein identification using mass spectrometry.** *Biochim Biophys Acta* 2003, **1646**:1-10.
9. Eng JK, McCormack AL, Yates JR, III: **An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.** *American Society for Mass Spectrometry* 1994, **5**:976-989.
10. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR: **Direct analysis of protein complexes using mass spectrometry.** *Nat Biotechnol* 1999, **17**:4385-4402.
11. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**:3551-3567.
12. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.** *Anal Chem* 2002, **74**:5383-5392.
13. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: **A statistical model for identifying proteins by tandem mass spectrometry.** *Anal Chem* 2003, **75**:4646-4658.
14. MacCoss MJ, Wu CC, Yates JR: **Probability-based validation of protein identifications using a modified SEQUEST algorithm.** *Anal Chem* 2002, **74**:5593-5599.
15. Eddes JS, Kapp EA, Frecklington DF, Connolly LM, Layton MJ, Moritz RL, Simpson RJ: **CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies.** *Proteomics* 2002, **2**:1097-1103.
16. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J: **OLAV: towards high-throughput tandem mass spectrometry data identification.** *Proteomics* 2003, **3**:1454-1463.
17. Magnin J, Masselot A, Menzel C, Colinge J: **OLAV-PMF: a novel scoring scheme for high-throughput peptide mass fingerprinting.** *J Proteome Res* 2004, **3**:55-60.

18. Colinge J, Masselot A, Cusin I, Mahe E, Niknejad A, Argoud-Puy G, Reffas S, Bederr N, Gleizes A, Rey PA et al.: **High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics.** *Proteomics* 2004, **4**:1977-1984.
19. Fenyo D, Beavis RC: **A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes.** *Anal Chem* 2003, **75**:768-774.
20. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra.** *Bioinformatics* 2004, **20**:1466-1467.
21. Eriksson J, Fenyo D: **Probioty: a protein identification algorithm with accurate assignment of the statistical significance of the results.** *J Proteome Res* 2004, **3**:32-36.
22. Craig R, Beavis RC: **A method for reducing the time required to match protein sequences with tandem mass spectra.** *Rapid Commun Mass Spectrom* 2003, **17**:2310-2316.
23. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm.** *J Proteome Res* 2004, **3**:958-964.
24. Field HI, Fenyo D, Beavis RC: **RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database.** *Proteomics* 2002, **2**:36-47.
25. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS et al.: **An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.** *Proteomics* 2005, **5**:3475-3490.
26. Li XJ, Zhang H, Ranish JA, Aebersold R: **Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry.** *Anal Chem* 2003, **75**:6648-6657.
27. Han DK, Eng J, Zhou H, Aebersold R: **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry.** *Nat Biotechnol* 2001, **19**:946-951.
28. Li XJ, Pedrioli PG, Eng J, Martin D, Yi EC, Lee H, Aebersold R: **A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry.** *Anal Chem* 2004, **76**:3856-3860.
29. **MSQuant.** <http://msquant.sourceforge.net/>. Last Visit 26-2-2007.
30. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC et al.: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
31. Maurer M, Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, Prokesch A, Scheideler M, Trajanoski Z: **MARS: microarray analysis, retrieval, and storage system.** *BMC Bioinformatics* 2005, **6**:101.
32. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3**:SOFTWARE0003.
33. **Human Proteome Organisation (HUPO).** <http://www.hupo.org>. Last Visit 21-9-2006.
34. Orchard S, Hermjakob H, Apweiler R: **The proteomics standards initiative.** *Proteomics* 2003, **3**:1374-1376.
35. **Protomics Standards Initiative (PSI).** <http://psidev.sourceforge.net>. Last Visit 3-10-2006.

36. Orchard S, Kersey P, Hermjakob H, Apweiler R: **Meeting Review: The HUPO Proteomics Standards Initiative meeting: towards common standards for exchanging proteomics data - Hinxton, Cambridge, UK, 19-20 October 2002.** *Comparative and Functional Genomics* 2003, **4**:16-19.
37. Orchard S, Kersey P, Zhu WM, Montecchi-Palazzi L, Hermjakob H, Apweiler R: **Meeting review: Progress in establishing common standards for exchanging proteomics data: the second meeting of the HUPO proteomics standards initiative.** *Comparative and Functional Genomics* 2003, **4**:203-206.
38. Orchard S, Zhu W, Julian RK, Jr., Hermjakob H, Apweiler R: **Further advances in the development of a data interchange standard for proteomics data.** *Proteomics* 2003, **3**:2065-2066.
39. Orchard S, Taylor CF, Hermjakob H, Weimin Z, Julian RKJ, Apweiler R: **Advances in the development of common interchange standards for proteomic data.** *Proteomics* 2004, **4**:2363-2365.
40. Orchard S, Hermjakob H, Julian RKJ, Runte K, Sherman D, Wojcik J, Zhu W, Apweiler R: **Common interchange standards for proteomics data: Public availability of tools and schema.** *Proteomics* 2004, **4**:490-491.
41. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C et al.: **The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data.** *Nat Biotechnol* 2004, **22**:177-183.
42. Orchard S, Hermjakob H, Binz PA, Hoogland C, Taylor CF, Zhu W, Julian RK, Jr., Apweiler R: **Further steps towards data standardisation: the Proteomic Standards Initiative HUPO 3(rd) annual congress, Beijing 25-27(th) October, 2004.** *Proteomics* 2005, **5**:337-339.
43. Orchard S, Montecchi-Palazzi L, Hermjakob H, Apweiler R: **The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments.** *Pac Symp Biocomput* 2005, 186-196.
44. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I et al.: **A systematic approach to modeling, capturing, and disseminating proteomics experimental data.** *Nat Biotechnol* 2003, **21**:247-254.
45. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: The proteomics identifications database.** *Proteomics* 2005, **5**:4046.
46. Keller A, Eng J, Zhang N, Li XJ, Aebersold R: **A uniform proteomics MS/MS analysis platform utilizing open XML file formats.** *Mol Sys Biology* 2005, E1-E8.
47. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *J Proteome Res* 2004, **3**:1234-1242.
48. Matthiesen R, Bunkenborg J, Stensballe A, Jensen ON, Welinder KG, Bauw G: **Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V20.** *Proteomics* 2004, **4**:2583-2593.
49. Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON: **VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins.** *J Proteome Res* 2005, **4**:2338-2347.
50. Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, Hussey P, Igra M, Maclean B, Lin CW, Detter A et al.: **Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments.** *J Proteome Res* 2006, **5**:112-121.
51. Wilke A, Ruckert C, Bartels D, Dondrup M, Goesmann A, Huser AT, Kespohl S, Linke B, Mahne M, McHardy A et al.: **Bioinformatics support for high-throughput proteomics.** *J Biotechnol* 2003, **106**:147-156.

52. Garden P, Alm R, Hakkinen J: **PROTEIOS: an open source proteomics initiative.** *Bioinformatics* 2005, **21**:2085-2087.
53. Shadforth I, Xu W, Crowther D, Bessant C: **GAPP: a fully automated software for the confident identification of human peptides from tandem mass spectra.** *J Proteome Res* 2006, **5**:2849-2852.
54. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas project.** *Nucleic Acids Res* 2006, **34**:D655-D658.
55. Kristensen DB, Brond JC, Nielsen PA, Andersen JR, Sorensen OT, Jorgensen V, Budin K, Matthiesen J, Veno P, Jespersen HM et al.: **Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data.** *Mol Cell Proteomics* 2004, **3**:1023-1038.
56. Shinkawa T, Taoka M, Yamauchi Y, Ichimura T, Kaji H, Takahashi N, Isobe T: **STEM: a software tool for large-scale proteomic data analyses.** *J Proteome Res* 2005, **4**:1826-1831.
57. Kohlbacher O, Reinert K, Gropl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M: **TOPP--the OpenMS proteomics pipeline.** *Bioinformatics* 2007, **23**:e191-e197.
58. **Proxeon.** <http://www.proxeon.com>. Last Visit 5-3-2007.
59. **Phenyx.** <http://www.phenyx-ms.com/>. Last Visit 5-3-2007.
60. **Scaffold.** <http://www.proteomesoftware.com/index.html>. Last Visit 5-3-2007.
61. **Agilent Technologies.** <http://www.chem.agilent.com/scripts/pds.asp?lpage=7771>. Last Visit 5-3-2007.
62. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nat Biotechnol* 1999, **17**:994-999.
63. Gygi SP, Rist B, Griffin TJ, Eng J, Aebersold R: **Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags.** *J Proteome Res* 2002, **1**:47-54.
64. Posewitz MC, Tempst P: **Immobilized gallium(III) affinity chromatography of phosphopeptides.** *Anal Chem* 1999, **71**:2883-2892.
65. Gruhler A, Olsen JV, Mohammed S, Mortensen P, Faergeman NJ, Mann M, Jensen ON: **Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway.** *Mol Cell Proteomics* 2005, **4**:310-327.
66. Wiese S, Reidegeld KA, Meyer HE, Warscheid B: **Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research.** *Proteomics* 2006.
67. Washburn MP, Wolters D, Yates JR, II I: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19**:242-247.
68. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A: **The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data.** *Mol Cell Proteomics* 2004, **3**:531-533.
69. Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S et al.: **PEDRo: a database for storing, searching and disseminating experimental proteomics data.** *BMC Genomics* 2004, **5**:68.
70. Quackenbush J: **Data standards for 'omic' science.** *Nat Biotechnol* 2004, **22**:613-614.
71. Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK, Jr., Apweiler R: **Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005).** *Proteomics* 2005, **5**:3552-3555.

72. Orchard S, Hermjakob H, Taylor C, Binz PA, Hoogland C, Julian R, Garavelli JS, Aebersold R, Apweiler R: **Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4-6, 2005.** *Proteomics* 2006.
73. **MIAPE.** <http://psidev.sourceforge.net/miape>. Last Visit 9-8-2006.
74. **MIAPE GE v1.0.** http://psidev.sourceforge.net/miape/MIAPE_GE_1.0.pdf. Last Visit 6-3-2007.
75. **MIAPE MSI v0.7.** http://psidev.sourceforge.net/miape/MIAPE_MSI_0.7.pdf. Last Visit 6-3-2007.
76. **MIAPE MS v2.1.** http://psidev.sourceforge.net/miape/MIAPE_MS_2.1.pdf. Last Visit 6-3-2007.
77. **MIAPE CC v0.1.** http://www.cs.man.ac.uk/~norm/MIAPE/MIAPE_SP_0.2.doc. Last Visit 3-9-2006.
78. **MIAPE CE v0.4.** http://psidev.sourceforge.net/miape/MIAPE_CE_0.4.pdf. Last Visit 6-3-2007.
79. **European Bioinformatics Institute (EBI).** <http://www.ebi.ac.uk>. Last Visit 6-3-2007.
80. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: The proteomics identifications database (vol. 5, Issue 13, pp. 3537-3545).** *Proteomics* 2005, 5:4046.
81. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide identifications for the proteomics community.** *Nucleic Acids Res* 2006, 34:D659-D663.
82. Armstrong E., Ball J., Bodoff S., Carson D.B., Evans I., Green D., Haase K., Jendrock E.: *The J2EE 1.4 Tutorial.* 8.2 edition 2005.
83. **Java Beans.** <http://java.sun.com/products/javabeans/>. Last Visit 5-3-2007.
84. Roman E: *Mastering Enterprise JavaBeans.* Wiley Computer Publishing; 2nd edition 2002.
85. **J2EE Java Message Service API Overview.** <http://java.sun.com/products/jms/overview.html>. Last Visit 5-3-2007.
86. Bien A: *J2EE Patterns. Entwurfsmuster fuer J2EE.* Addison Wesley;2003.
87. **Core J2EE Patterns - Session Facade.** <http://java.sun.com/blueprints/corej2eepatterns/Patterns/SessionFacade.html>. Last Visit 5-3-2007.
88. Brown S, Burdick R, Falkner J, Galbraith B, Johnson R, Kim L, Kochmer C, Kristmundsson T, Li S, Malks D et al.: *Professional JSP.* Wrox Press; 2nd edition 2001.
89. **VelocityStruts User Guide.** <http://velocity.apache.org/tools/devel/struts/userguide.html>. Last Visit 5-3-2007.
90. **The Object Management Group.** <http://www.omg.org/>. Last Visit 5-3-2007.
91. **OMG Model Driven Architecture.** <http://www.omg.org/mda/>. Last Visit 6-2-2007.
92. *Model Driven Architecture - A technical perspective.* 2001. [<ftp://ftp.omg.org/pub/docs/ab/01-02-04.pdf>]
93. Kleppe A, Warmer J.B., Bast W, Watson A: *MDA Explained - The Model Driven Architecture: Practice and Promise.* Addison-Wesley;2003.
94. **JSR-000244 Java™ Platform, Enterprise Edition 5 Specification.** <http://jcp.org/aboutJava/communityprocess/pfd/jsr244/>. Last Visit 5-3-2007.
95. **.NET.** <http://www.microsoft.com/net/>. Last Visit 5-3-2007.

96. Bohlen M: **QVT und Multi-Metamodell-Transformationen in MDA**. *OBJECTspectrum* 2006, **02**:51-57.
97. Sendall S, Kozaczynski W: **Model Transformation: The Heart and Soul of the Model-Driven Software Development**. *IEEE Software* 2003, **20**:42-45.
98. Ignjatovic M: **"Query/View/Transformation": Ein neuer OMG-Standard**. *OBJECTspectrum* 2006, **02**:46-50.
99. **AndroMDA**. <http://www.andromda.org>. Last Visit 10-11-2006.
100. **MagicDraw**. <http://www.magicdraw.com/>. Last Visit 22-2-2007.
101. **Poseidon**. <http://www.gentleware.com>. Last Visit 5-3-2007.
102. **Ant**. <http://ant.apache.org>. Last Visit 5-3-2007.
103. **Velocity**. <http://velocity.apache.org>. Last Visit 9-2-2007.
104. Friese P, Bohlen M: **Erfolgreicher Einsatz von AndroMDA bei Lufthansa Systems**. *OBJECTspectrum* 2006, **03**:62-68.
105. **XDoclet**. <http://xdoclet.sourceforge.net>. Last Visit 5-3-2007.
106. **JBoss**. <http://jboss.com>. Last Visit 7-3-2007.
107. Hackl H, Maurer M, Mlecnik B, Hartler J, Stocker G, Miranda-Saavedra D, Trajanoski Z: **GOLDDb: Genomics of lipid-associated disorders database**. *BMC Genomics* 2004, **5**:93.
108. Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z: **PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways**. *Nucleic Acids Res* 2005, **33**:W633-W637.
109. Codd E.F.: *The relational model for database management*. Boston MA: Addison-Wesley; 2 edition 1990.
110. **MSSQL**. <http://www.microsoft.com/sql/default.mspx>. Last Visit 5-3-2007.
111. **MySQL**. <http://www.mysql.com>. Last Visit 5-3-2007.
112. **PostgreSQL**. <http://www.postgresql.org>. Last Visit 5-3-2007.
113. **Oracle**. <http://www.oracle.com>. Last Visit 5-3-2007.
114. **Enterprise JavaBeansQuery Language**. http://java.sun.com/j2ee/tutorial/1_3-fcs/doc/EJBQL.html. Last Visit 5-3-2007.
115. **JDBC**. <http://java.sun.com/javase/technologies/database>. Last Visit 5-3-2007.
116. Stocker G: **Computational Environment for Cellular Imaging by Fluorescence Microscopy**. *PhD Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2007.
117. **SOAP**. <http://www.w3.org/TR/soap/>. Last Visit 5-3-2007.
118. Zeller D.: **Design and Development of a User Management System for Molecular Biology Database System**. *Master Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2003.
119. **IMP Protein Chemistry Facility**. http://www.imp.ac.at/protein/pro_hp.html. Last Visit 2007.
120. Molidor R: **Design and Development of a Bioinformatics Platform for Cancer Immunogenomics**. *PhD Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2004.

121. Truskaller, T.: **Data Integration into a Gene Expression Database**. *Master Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2003.
122. Diffie W, Hellman ME: **New Directions in Cryptography**. *IEEE Trans on Info Theory* 1976, **IT-22**:644-654.
123. Thallinger GG, Baumgartner K, Pirklbauer M, Uray M, Pauritsch E, Mehes G, Buck CR, Zatloukal K, Trajanoski Z: **TAMEE: data management and analysis for tissue microarrays**. *BMC Bioinformatics* 2007, **8**:81.
124. Rader R: **Design and Development of a Database for Protein-Protein Interaction in Crystals**. *Master Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2005.
125. Pabinger S: **Development of a Web-based application for managing and analyzing real-time PCR experiments**. *Master Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2006.
126. Achuthan P.: **StocksDB: Design and Development of a Database Application for the Management of Biological Stocks**. *Master Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2004.
127. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R et al.: **A common open representation of mass spectrometry data and its application to proteomics research**. *Nat Biotechnol* 2004, **22**:1459-1466.
128. **National Center for Biotechnology Information**. <http://www.ncbi.nlm.nih.gov/>. Last Visit 20-4-2007.
129. **MSDB**. <http://csc-fserve.hh.med.ic.ac.uk/msdb.html>. Last Visit 2006.
130. **Java Regular Expressions**. <http://java.sun.com/j2se/1.5.0/docs/api/java/util/regex/Pattern.html>. Last Visit 6-4-2007.
131. Kislinger T, Rahman K, Radulovic D, Cox B, Rossant J, Emili A: **PRISM, a generic large scale proteomic investigation strategy for mammals**. *Mol Cell Proteomics* 2003, **2**:96-106.
132. Sturn A: **Comparative Analysis of Human and Mouse Transcriptomes**. *PhD Thesis*. Institute for Genomics and Bioinformatics, TU Graz; 2005.
133. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res* 2002, **30**:1575-1584.
134. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.
135. Savitzky A, Golay MJE: **Smoothing and Differentiation of Data by Simplified Least Squares Procedures**. *Analytical Chemistry* 1964, **36**:1627-1639.
136. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor**. *Bioinformatics* 2004, **20**:426-427.
137. Roepstorff P, Fohlman J: **Proposal for a common nomenclature for sequence ions in mass spectra of peptides**. *Biomed Mass Spectrom* 1984, **11**:601.
138. Johnson RS, Martin SA, Biemann K, Stults JT, Watson JT: **Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine**. *Anal Chem* 1987, **59**:2621-2625.
139. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT et al.: **Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling**. *Cell* 2006, **125**:173-186.

140. Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data.** *Bioinformatics* 2002, **18**:207-208.
141. Yu JS, Ongarello S, Fiedler R, Chen XW, Toffolo G, Cobelli C, Trajanoski Z: **Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data.** *Bioinformatics* 2005, **21**:2200-2209.
142. Wulfkuhle JD, Liotta LA, Petricoin EF: **Proteomic applications for the early detection of cancer.** *Nat Rev Cancer* 2003, **3**:267-275.
143. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC et al.: **Use of proteomic patterns in serum to identify ovarian cancer.** *Lancet* 2002, **359**:572-577.
144. Wulfkuhle JD, McLean KC, Paweletz CP, Sgroi DC, Trock BJ, Steeg PS, Petricoin EF: **New approaches to proteomic analysis of breast cancer.** *Proteomics* 2001, **1**:1205-1215.
145. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z et al.: **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men.** *Cancer Res* 2002, **62**:3609-3614.
146. Cazares LH, Adam BL, Ward MD, Nasim S, Schellhammer PF, Semmes OJ, Wright GLJ: **Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry.** *Clin Cancer Res* 2002, **8**:2541-2552.
147. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R, Petricoin EF: **Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence.** *Environ Health Perspect* 2004, **112**:1622-1627.
148. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GLJ, Qu Y, Potter JD, Winget M, Thornquist M, Feng Z: **A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection.** *Biostatistics* 2003, **4**:449-463.
149. Baggerly KA, Morris JS, Coombes KR: **Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.** *Bioinformatics* 2004, **20**:777-785.
150. Courchesne PL, Luethy R, Patterson SD: **Comparison of in-gel and on-membrane digestion methods at low to sub-pmol level for subsequent peptide and fragment-ion mass analysis using matrix-assisted laser-desorption/ionization mass spectrometry.** *Electrophoresis* 1997, **18**:369-381.
151. Davis MT, Lee TD: **Rapid protein identification using a microscale electrospray LC/MS system on an ion trap mass spectrometer.** *J Am Soc Mass Spectrom* 1998, **9**:194-201.
152. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR, III: **Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level.** *Anal Chem* 1997, **69**:767-776.
153. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.** *Proc Natl Acad Sci U S A* 2000, **97**:9390-9395.
154. Fenn J.B.: **Electrospray Ionization Mass Spectrometry: How It All Began.** *Journal of Biomolecular Techniques* 2002, **13**:101-118.
155. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules.** *Science* 1989, **246**:64-71.
156. Lim MS, Elenitoba-Johnson KS: **Proteomics in pathology research.** *Lab Invest* 2004, **84**:1227-1244.

157. Glish G.L: **Multiple stage mass spectrometry: the next generation tandem mass spectrometry experiment.** *Analyst* 2006, **119**:533-537.
158. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C: **Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.** *Proc Natl Acad Sci U S A* 1993, **90**:5011-5015.
159. James P, Quadroni M, Carafoli E, Gonnet G: **Protein identification by mass profile fingerprinting.** *Biochem Biophys Res Commun* 1993, **195**:58-64.
160. Yates JR, III, Speicher S, Griffin PR, Hunkapiller T: **Peptide mass maps: a highly informative approach to protein identification.** *Anal Biochem* 1993, **214**:397-408.
161. Pappin DJ, Hojrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass fingerprinting.** *Curr Biol* 1993, **3**:327-332.
162. Yates JR, III, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.** *Anal Chem* 1995, **67**:1426-1436.
163. Clauser KR, Baker P, Burlingame AL: **Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching.** *Anal Chem* 1999, **71**:2871-2882.
164. Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information.** *Anal Chem* 2000, **72**:2482-2489.
165. **PepMapper.** <http://wolf.bms.umist.ac.uk/mapper/>. Last Visit 2006.
166. **Aldente.** <http://www.expasy.org/tools/aldente/>. Last Visit 11-9-2006.
167. Fenyo D, Qin J, Chait BT: **Protein identification using mass spectrometric information.** *Electrophoresis* 1998, **19**:998-1005.
168. **SONAR.** <http://65.219.84.5/service/prowl/sonar.html>. Last Visit 2006.

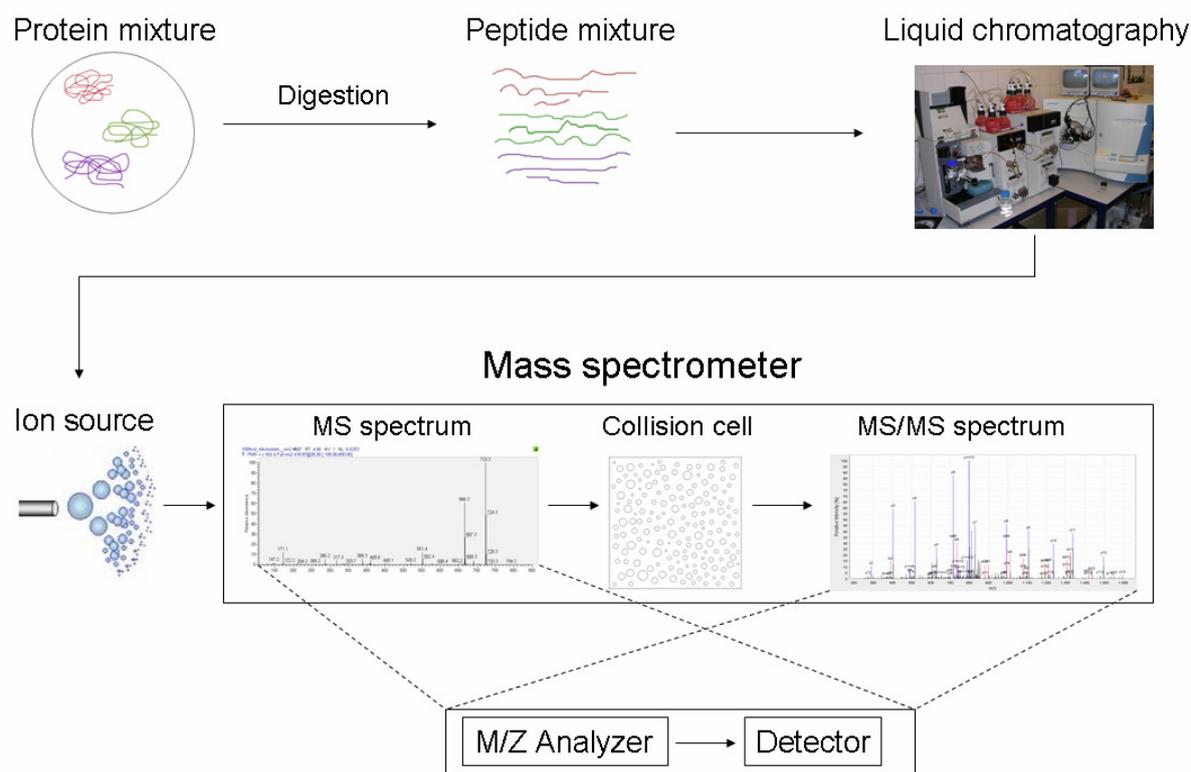
Appendix A – Mass Spectrometry

In MS-based proteomics two approaches for the research of complex diseases like cancer emerged. The so-called ‘expression-profiling’ focuses on the development of markers for the recognition of (early-stage) cancer by screening methods. The idea behind it is to find discriminatory differences by the comparison of spectra of healthy patients with spectra of patients who are carrying the disease. Several studies showed quite promising results concerning sensitivity and specificity [141-148]. Many of these studies were conducted using the surface enhanced laser desorption ionization (SELDI). This approach helps to distinguish between cancer and healthy patients but gives no insights in the proteins involved in the disease development. Additionally, reproducibility problems with the SELDI approach have been observed [149]. The second approach tries to identify the proteins in complex mixtures and their functions in complex diseases. It is separated into three steps: first, the characterization of proteins and their post-translational modifications; second, the differential comparison of protein levels; third, analysis of protein-protein interactions [1]. For this approach liquid chromatography tandem mass spectrometry (LC-MS/MS) is one of the most commonly used techniques (see figure “Schematic overview of a typical LC-MS/MS experiment”).

A MS based experiment starts with the sequential application of several separation techniques to extract the desired protein mixture, using 1D-gels, 2D-gels, column chromatography, and/or chemical separation techniques. The protein mixture is digested enzymatically, usually by trypsin. This step is usually already done while the desired proteins are extracted from the whole mixture with in-gel digestion [150].

The peptide mixture is solubilized and enters a liquid chromatography (LC) column. The peptides are separated by a certain physical or chemical property, depending on the protein species of interest. The most common methods in liquid chromatography are: ion exchange chromatography (charge), ion exclusion chromatography (size), reversed phase chromatography (hydrophobicity), affinity chromatography (biological interactions). For proteomics analysis microcapillary column high performance liquid column chromatography (HPLC), directly coupled to the mass spectrometer is the method of choice [10,151-153]. In the reverse phase approach two solutions with completely different pH-values are mixed at different rates to produce a gradient over time. The peptide mixture is washed with the induced gradient, and peptides adherent to this pH-value are swept out of the mixture carried along to a nano-spray needle. Between the nano-spray needle and the inlet to the mass spectrometer a high voltage (3-4 kV) is applied. As a result the sample emerging from the tip is dispersed into an

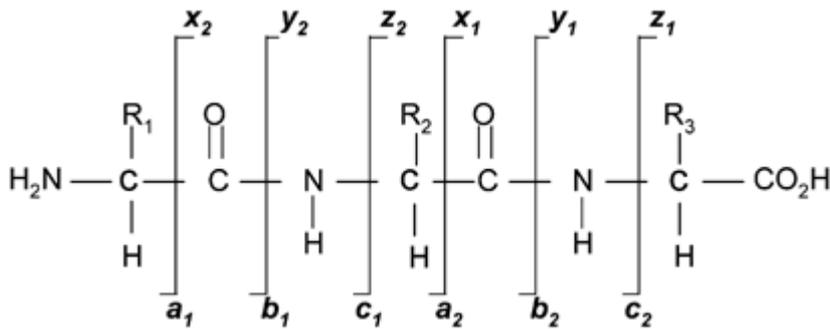
aerosol of highly charged droplets. The droplets diminish in size and the electric charge density on their surfaces increases. The repulsion between the charges on the surface reach a higher extend than the surface tensions. The ions leave the droplet through what is known as a “Taylor cone”. The charged droplets decrease in size by solvent evaporation. At the end of this process charged sample ions are released from the droplets, free of solvent. The ions pass through a small aperture into the analyzer of the mass spectrometer, which is held under high vacuum [154,155]. The described method is used for all of the performed experiments and is called electrospray ionization (ESI).



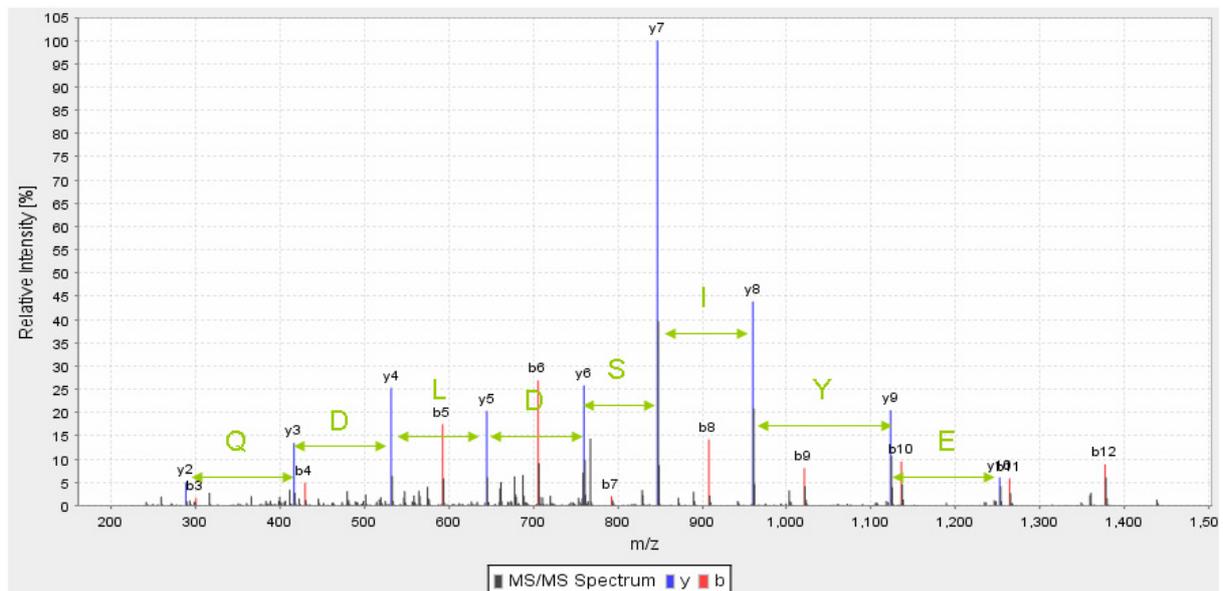
Schematic overview of a typical LC-MS/MS experiment. The complex protein mixture, which has to be analyzed, is digested enzymatically. The solubilized peptides are then separated by liquid chromatography. At the end of the column they are ionized and enter the mass spectrometer via a nano-spray needle. In a first run the peptides are separated according to their mass to charge ratios (m/z). The particles which caused the most abundant peaks are reselected by the mass spectrometer automatically and guided against a collision cell. The peptides fragment at the peptide backbone, and fragment ion series are generated. In a second run a mass spectrum of the fragments is measured. This step can be repeated several times. A run itself consists of two steps: the ions are passing through the mass analyzer and they are collected by a detector.

In the mass spectrometer the ions are separated according to their mass to charge ratio (m/z) by an m/z analyzer. Afterwards the signal intensities caused by the ions are collected by a detector. The resulting mass spectrum gives information about the mass of the intact peptide [4]. Due to the fact that different peptides can have the same or similar masses (e.g. same amino acids but different sequence) the results of just one MS run can cause ambiguities. Therefore, the most abundant ions (the ones with the highest intensities) are reselected by the mass spectrometer and guided against a collision cell. The peptides break predominantly along their backbone and the results are different ion series (see

figure “Peptide ion fragmentation”). The fragments are measured again by the mass spectrometer. The main benefit of this method is that the differences between the peaks of the fragments correspond to difference in mass of one amino acid (see figure “MS/MS spectrum of the peptide IGEEYISDLDQLR”).



Peptide ion fragmentation (taken from [156]). The $\text{R}_{1,2,3}$ correspond to the amino acid residues. The vertical bars which are labeled with a, b, c, x, y, and z are the possible breaking points of the bonds. E.g. b₁ and y₂ have the same breaking bond. b₁ would correspond to the fragment starting at the left side (1 amino acid) and y₂ would correspond to fragment starting from the right side. This nomenclature has been first proposed by Roepstorff [137] and adapted by Johnson [138]. The weakest bond is between carbonyl group and the amide nitrogen group, so the most commonly observed fragments are coming from the b and y series.



MS/MS spectrum of the peptide IGEEYISDLDQLR. The distance between the y-fragments corresponds to mass of the corresponding amino acid. The same could be done for the b-series.

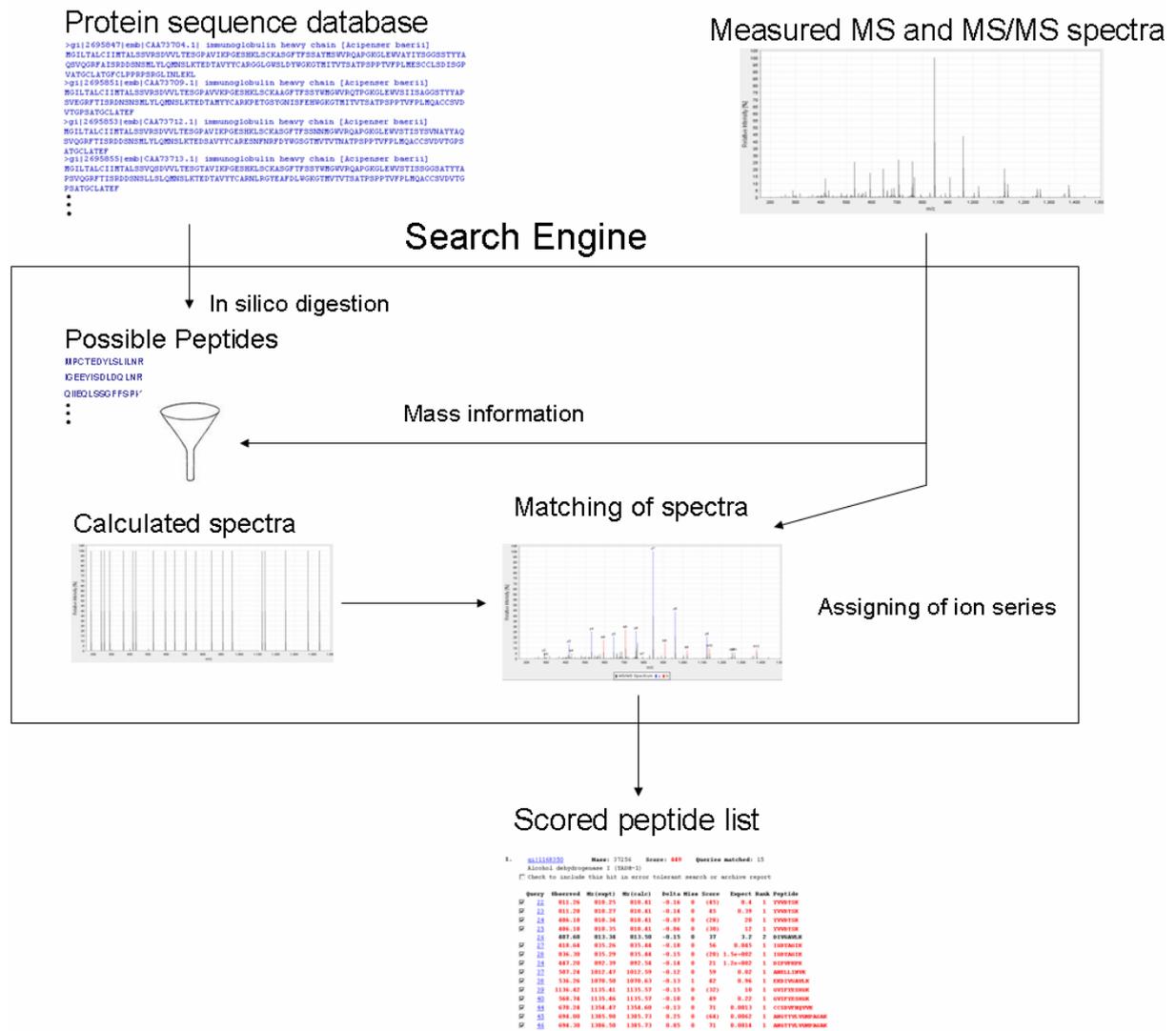
The MS spectrum combined with the tandem mass spectrum (MS/MS spectrum) gives information about the total mass of the peptide as well as information about the sequence of the peptide. It is also possible that the bonds of post translational modifications break rather than the peptide backbone (because it is weaker). Thus, in the MS/MS solely the precursor peak with a mass loss of the post

translational modification is observed and the peptide fragments are hardly detectable. Therefore, the fragmentation and detection of a MS/MS spectrum can be repeated several times [157]. Afterwards, the ion mass of the analyzed peptide is put on an exclusion list for a configurable period of time that less abundant peptides can be detected easier.

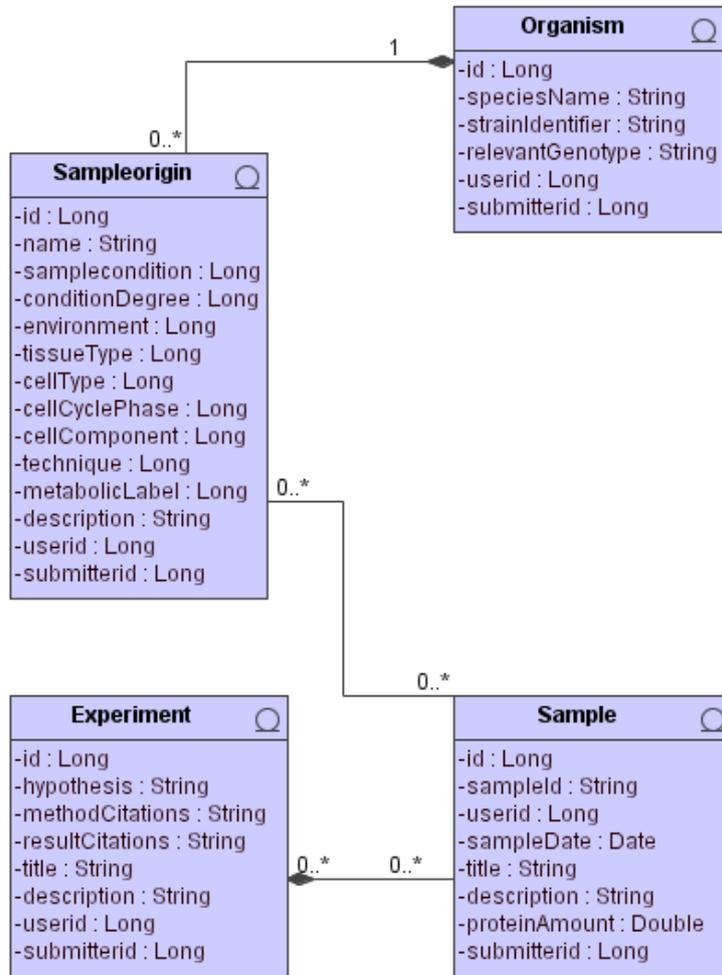
Appendix B – Database search engines

The manual evaluation a mass spectrometry experiment is not feasible because the results exceed several 100MB and millions of spectra. Therefore, the assignment to real world peptides is done in an automated manner by database search engines.

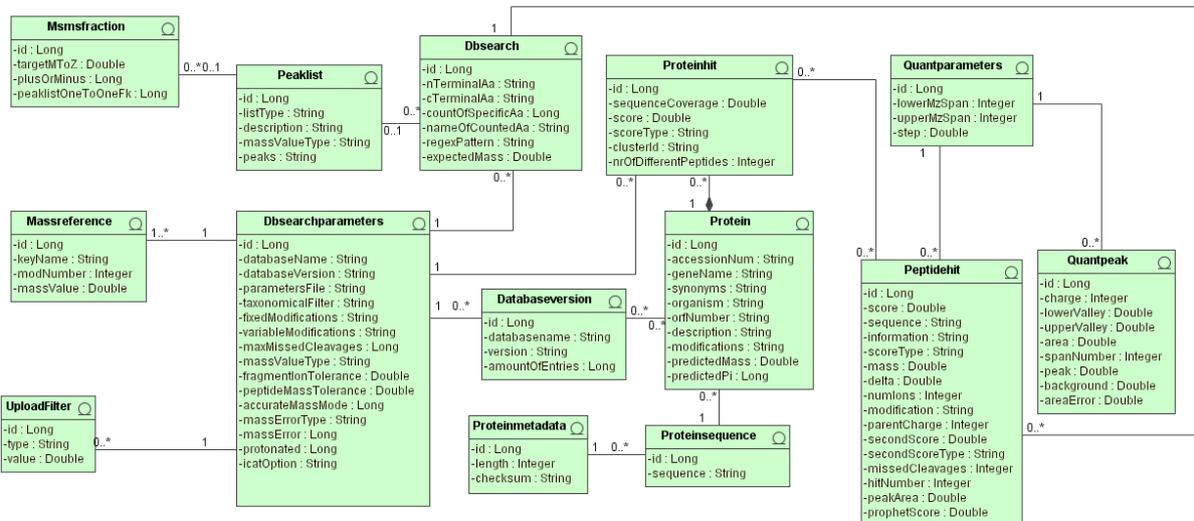
All commonly used search engines base on the same principle: they are using a protein sequence database, digest the protein sequences in silico with the same enzyme used for the mass spectrometry experiment, and try to assign peptides to the experimentally obtained mass spectra. Two common strategies are used [6]: peptide mass fingerprinting (PMF) and peptide fragmentation fingerprinting (PFF). In the figure “MS/MS spectrum of the peptide IGEEYISDLDQLR” the input and the output could be the same for both of them. PMF uses solely the mass information obtained from the MS spectra. The in silico obtained masses of the peptides are used as mass map and the experimentally obtained masses are correlated to them [158-161]. Reliable results depend strongly on the resolution and accuracy of the mass spectrometer. Furthermore, different peptide sequences of a protein database could match the observed mass randomly. PFF goes a step further (see figure “Schematic overview of the function of a PFF search engine”). In silico obtained masses are used just as filter. For the remaining peptide sequences a presumable MS/MS spectrum (for the expected ion series) is calculated and compared to the corresponding measured MS/MS spectrum [7,162]. Since the fingerprints are compared, deviations in the mass accuracy or mass shifts do not have as grave effects as in PMF. The common search engines for PMF are MS-Fit [163], ProFound [164], PepMapper [165], and Aldente [166], and for PFF SEQUEST [9,10], Mascot [11], SpectrumMill [61], X!Tandem [20], OMSSA [23], PepFrag [167], SONAR [168], and Phenyx [59]. Some of the PFF engines provide the possibility to search PMF as well. Studies [6,25] showed that each of them has different properties concerning specificity and sensitivity, and a combination of different methods can be useful.



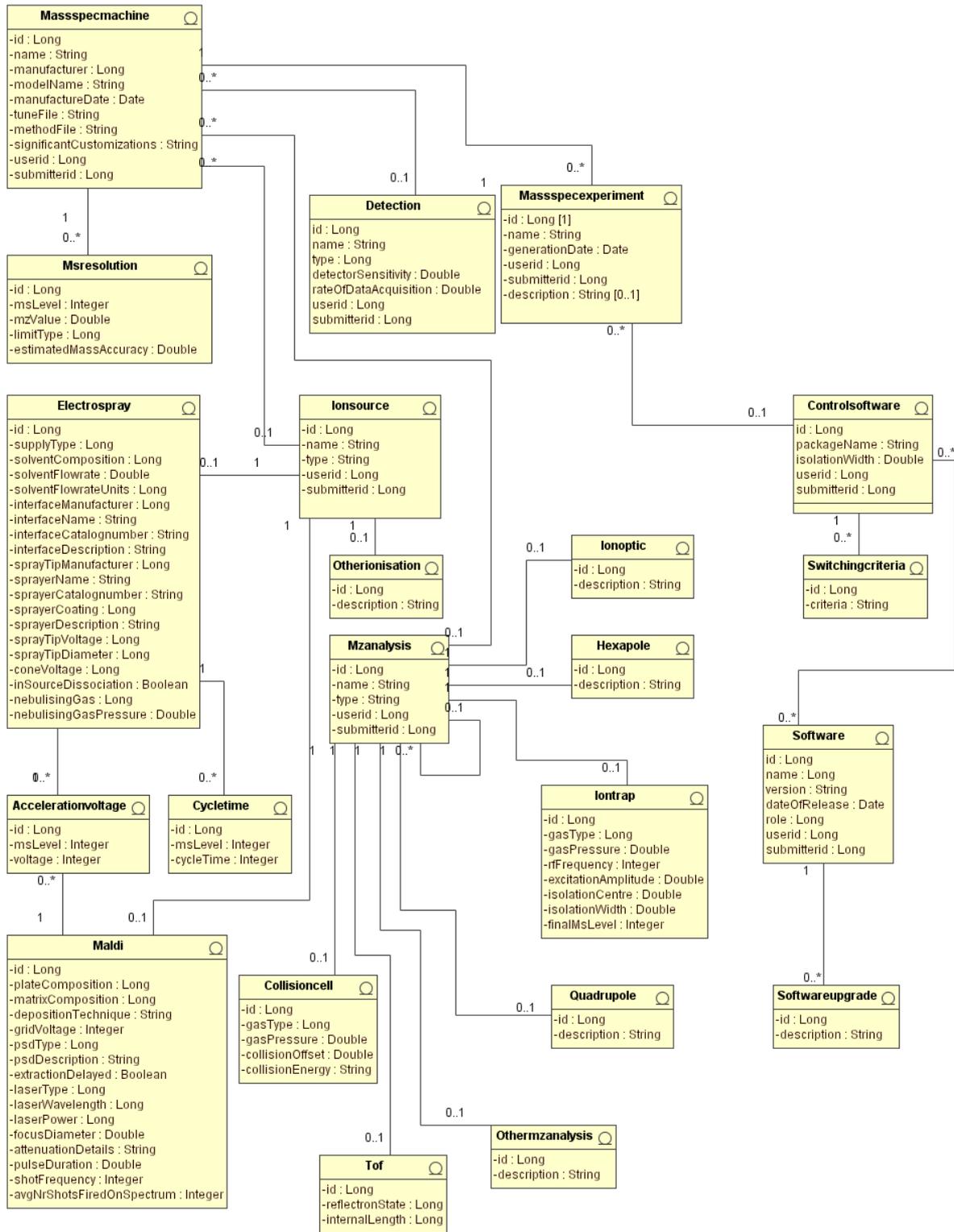
Schematic overview of the function of a PFF search engine. The PFF search engine requires a protein database and the measured MS and MS/MS spectra. The protein sequences in the database are digested in silico which leads to a list of peptides. The mass information from the MS spectrum is taken to filter the peptide candidate list. The search engine calculates ideal in silico spectra of the remaining candidate sequences. The in silico spectra are matched to the measured MS/MS spectrum and a score is assigned to a peptide hit. The output of the search engines results in a list of probable peptides (where the quality of the match is expressed by the score) and the corresponding proteins.



MASPECTRAS Sample Generation schema



MASPECTRAS MS informatics schema



MASPECTRAS Mass spectrometry schema

Glossary

<i>AAS</i>	Authentication and Authorization System
<i>ACL</i>	Access Control List
<i>API</i>	Application Interface
<i>BMP</i>	Bean Managed Persistence
<i>CASE</i>	Computer Aided Software Engineering
<i>CC</i>	Column Chromatography
<i>cDNA</i>	complementary DeoxyriboNucleic Acid
<i>CMP</i>	Container Managed Persistence
<i>CPAS</i>	Computational Portal and Analysis System
<i>DAO</i>	Data Access Object
<i>DBMS</i>	DataBase Management System
<i>DNA</i>	DeoxyriboNucleic Acid
<i>EAR</i>	Enterprise ARchive
<i>EBI</i>	European Bioinformatics Institute
<i>EIS</i>	Enterprise Information System
<i>EJB</i>	Enterprise Java Bean
<i>EJB-QL</i>	Enterprise Java Beans Query Language
<i>EPiR</i>	Experimental Peptide Identification Repository
<i>ESI</i>	Electro-Spray Ionization
<i>GAPP</i>	Genome Annotating Proteomic Pipeline
<i>GE</i>	Gel Electrophoresis
<i>GI</i>	Gel (image) Informatics
<i>GPM</i>	The Global Proteome Machine
<i>HPLC</i>	High-Performance Liquid Chromatography
<i>HTML</i>	Hyper Text Markup Language
<i>HTTP</i>	Hyper Text Transfer Protocol
<i>HTTPS</i>	Hyper Text Transfer Protocol Secure
<i>HUPO</i>	HUman Proteome Organization
<i>ICPL</i>	Isotope Coded Protein Label
<i>J2EE</i>	Java 2 Enterprise Edition

JAR	Java ARchive
JDBC	Java DataBase Connectivity
JMS	Java Messaging Service
JSP	Java Server Page
LC	Liquid Chromatography
LC-MS/MS	Liquid Chromatography Tandem Mass Spectrometry
MARS	Microarray Analysis and Retrieval System
MASPECTRAS	MAss SPECTRometry Analysis System
MDA	Model Driven Architecture
MIAME	MInimum Information About a Microarray Experiment
MIAPE	MInimum Information About a Proteomics Experiment
MOF	Meta Objects Facility
MS	Mass Spectrometry
MSDB	Mass Spectrometry protein sequence DataBase
MSI	Mass Spectrometry Informatics
MS/MS	Tandem Mass Spectrometry
MVC	Model View Controller
m/z	mass to charge ratio
NCBI	National Center for Biotechnology Information
nanoRP-HPLC	nano Reverse Phase – High-Performance Liquid Chromatography
nr	non-redundant
OMG	Object Management Group
OMSSA	Open Mass Spectrometry Search Algorithm
PCR	Polymerase Chain Reaction
PEDRo	Proteome Experiment Data Repository
PFF	Peptide Fragmentation Fingerprinting
PIM	Platform Independent Model
PMF	Peptide Mass Fingerprinting
PRIDE	PRoteomics IDentifications database
PSI	Proteomics Standards Initiative
PSM	Platform Specific Model
PTM	Post Translational Modification
RDBMS	Relational DataBase Management System
RP-HPLC	Reverse Phase High-Performance Liquid Chromatography
RTPCR	Real-Time Polymerase Chain Reaction
SELDI	Surface Enhanced Laser Desorption Ionization
SOAP	Simple Object Access Protocol

<i>SQL</i>	Structured Query Language
<i>STEM</i>	STrategic Extractor for Mascot's results
<i>TOPP</i>	The Open MS Proteomics Pipeline
<i>TPP</i>	Trans-Proteomic Pipeline
<i>TAMEE</i>	Tissue Array Management and Evaluation Environment
<i>UML</i>	Unified Modeling Language
<i>VEMS</i>	Virtual Expert Mass Spectromist
<i>WAR</i>	Web ARchive
<i>XMI</i>	XML Metadata Interchange
<i>XML</i>	eXtensible Markup Language

Acknowledgements

This work has been supported by the Austrian Academy of Sciences, the Bioinformatics Integration Network (BIN II) of the Austrian Genome Research Program (GENAU), and the Christian Doppler Laboratory in collaboration with the Austrian Research Centers.

I would like to express my deepest gratitude to my mentor Zlatko Trajanoski for his encouragements and belief in me. I am deeply grateful to Karl Mechtler who gave me the possibility to gain insights into the world of mass spectrometry, and teaching me the peculiarities of this field. I want to thank the members of his group, Andreas Schmidt and Christoph Stingl, and the members of the Biocenter in Innsbruck, Stefan Ascher and Sandra Morandell for their fruitful discussions.

Further thanks to the members of the “Bioinformatics group” and those at the “Institute for Genomics and Bioinformatics” for their contributions, support and friendship. Thanks to the people who did excellent preliminary work in their applications that such a rapid progress has been made possible (in alphabetical order): Premanand Achuthan, Michael Maurer, Bernhard Mlecnik, Robert Molidor, and Thomas Truskaller. Special thanks to the people who have directly contributed to this work: Thomas Burkard, Thomas Fuchs, Erik Körner, Robert Rader, Andreas Scheucher, Gernot Stocker, Alexander Sturn, and Gerhard Thallinger.

I am indebted to the members of my family for their unfailing support and my companion in life Birgit, for accompanying me, her love, encouragement and support.

Publications

Journals

Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, Körner E, Rader R, Schmidt A, Mechtler K, and Trajanoski Z: **MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data.** *submitted to BMC Bioinformatics*

Maurer M, Molitor R, Sturn A, Hartler J, Hackl H, Stocker G, Prokesch A, Scheideler M, and Trajanoski Z: **MARS: microarray analysis, retrieval, and storage system.** *BMC Bioinformatics* 2005, **6**:101.

Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, and Trajanoski Z: **PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.** *Nucleic Acids Res* 2005, **33**:W633-W637.

Hackl H, Maurer M, Mlecnik B, Hartler J, Stocker G, Miranda-Saavedra D, and Trajanoski Z: **GOLDdb: Genomics of lipid-associated disorders database.** *BMC Genomics* 2004, **5**:93.

MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data

Jürgen Hartler^{*§}, Gerhard G. Thallinger^{*}, Gernot Stocker^{*}, Alexander Sturm^{*}, Thomas R. Burkard^{*}, Erik Körner[†], Robert Rader^{*}, Andreas Schmidt[#], Karl Mechtler[‡], and Zlatko Trajanoski^{*}

^{*}Institute for Genomics and Bioinformatics and Christian-Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria.

[§]Austrian Research Centers GmbH –ARC, eHealth Systems, Reininghausstrasse 13/1, 8020 Graz, Austria

[†]FH Joanneum, Kapfenberg, Werk-VI-Straße 46, 8605 Kapfenberg, Austria.

[‡]Research Institute for Molecular Pathology, Dr. Bohr-Gasse 7, 1030 Vienna, Austria.

[#] Christian Doppler Laboratory for Proteome Analysis, Dr. Bohr-Gasse 3, 1030 Vienna, Austria

Correspondence: Zlatko Trajanoski

Institute for Genomics and Bioinformatics

Graz University of Technology

Phone: +43 316 873 5332

Fax: +43 316 873 5340

E-mail: zlatko.trajanoski@tugraz.at

Abstract

We have developed MAss SPECTRometry Analysis System (MASPECTRAS), a platform for management and analysis of proteomics LC-MS/MS data. MASPECTRAS is based on the Proteome Experimental Data Repository (PEDRo) relational database scheme and follows the guidelines of the Proteomics Standards Initiative (PSI). Analysis modules include: 1) import and parsing of the results from the search engines SEQUEST, Mascot, Spectrum Mill, X! Tandem, and OMSSA; 2) peptide validation, 3) clustering of proteins based on Markov Clustering and multiple alignments; and 4) quantification using the Automated Statistical Analysis of Protein Abundance Ratios algorithm (ASAPRatio). The system provides customizable data retrieval and visualization tools, as well as export to PRoteomics IDentifications public repository (PRIDE). MASPECTRAS is freely available at <http://genome.tugraz.at/MASPECTRAS>.

BACKGROUND

The advancement of genomic technologies – including microarray, proteomic and metabolic approaches – have led to a rapid increase in the number, size and rate at which genomic datasets are generated. Managing and extracting valuable information from such datasets requires the use of data management platforms and computational approaches. In contrast to genome sequencing projects, there is a need to store much more complex ancillary data than would be necessary for genome sequences. Particularly the need to clearly describe an experiment and report the variables necessary for data analysis became a new challenge for the laboratories. Furthermore, the vast quantity of data associated with a single experiment can become problematic at the point of publishing and disseminating results. Fortunately, the communities have recognized and tackled the problem through the development of standards for the capturing and sharing of experimental data. The microarray community arranged to define the critical information necessary to effectively analyze a microarray experiment and developed the Minimal Information About a Microarray Experiment (MIAME)[1] . Subsequently, MIAME was adopted by scientific journals as a prerequisite for publications and several software platforms supporting MIAME were developed [2,3] .

The principles underlying MIAME have reasoned beyond the microarray community. The Proteomics Standard Initiative (PSI) [4] aims to define standards for data representation in proteomics analogous to that of MIAME and developed Minimum Information About a Proteomics Experiment (MIAPE) [5]. An implementation independent approach for defining the data structure of a Proteomics Experiment Data repository (PEDRo) [6] was developed using unified modeling language (UML) and a PSI compliant public repository was set up [7]. Hence, given the defined standards and available public repositories proteomics laboratories computational systems can now be developed to support proteomics laboratories and enhance data dissemination.

To meet the needs for high-throughput MS laboratories several tools and platforms covering various parts of the analytical pipeline were recently developed including the Trans Proteomics Pipeline [8],

The Global Proteome Machine [9], VEMS [10,11], CPAS [12], CHOMPER [13], ProDB [14], PROTEIOS [15], GAPP [16], PeptideAtlas [17], EPIR [18], STEM [19], and TOPP [20] (see table 1 for a comparison of the features). However, to the best of our knowledge there is currently no academic or commercial data management platform supporting MIAPE and enabling PRoteomics IDentifications database (PRIDE) export. Moreover, it became evident that several search engines should be used to validate proteomics results [21]. Hence, a system enabling comparison of the results generated by the different search engines would be of great benefit. Additionally, integration of algorithms for peptide validation, protein clustering and protein quantification into a single analytical pipeline would considerably facilitate analyses of the experimental data.

We have therefore developed the MAAss SPECTRometry Analysis System (MASPECTRAS), a web-based platform for management and analysis of proteomics liquid chromatography tandem mass spectrometry (LC-MS/MS) data supporting MIAPE. MASPECTRAS was developed using state-of-the-art software technology and enables data import from five common search engines. Analytical modules are provided along with visualization tools and PRIDE export as well as a module for distributing intensive calculations to a computing cluster.

ANALYSIS PIPELINE

MASPECTRAS extends the PEDRo relational database scheme and follows the guidelines of the PSI. It accepts the native file formats from SEQUEST [22], Mascot [23], Spectrum Mill [24], X! Tandem [25], and OMSSA [26]. The core of MASPECTRAS is formed by the MASPECTRAS analysis platform (Figure 1). The platform encompasses modules for the import and parsing data generated by the above mentioned search engines, peptide validation, protein clustering, protein quantification, and a set of visualization tools for post-processing and verification of the data, as well as PRIDE export.

Import and Parsing Data from Search Engines

There are several commercial and academic search engines for proteomics data. Based on known protein sequences stored in a database, these search engines perform *in silico* protein digestion to calculate theoretical spectra for the resulting peptides and compare them to the obtained ones. Based on the similarity of the two spectra, a probability score is assigned. The results (score, peptide sequence, etc.) are stored in a single or in multiple files, and often only an identification string for the protein is stored whereas the original sequence is discarded. However, the search engines are storing different identification strings for the proteins (e.g. X! Tandem: gil231300|pdb|8GPBI; Spectrum Mill: 231300). Moreover, several databases are not using common identifiers (eg. National Center for Biotechnology Information non redundant (NCBI nr): gil6323680; Mass Spectrometry protein sequence DataBase (MSDB) [27]: S39004). In order to compare the search results from different search engines additional information from the corresponding sequence databases is needed. The format of the accession string has to be known to retrieve the protein sequence and additional required information from the sequence database, like protein description, or the organism the protein belongs to. The only common basis within the different databases used by the search algorithms is the sequence information. In order to make results of different algorithms comparable and to find the corresponding proteins in the different result files the sequence information is taken as unique identification criteria.

We have developed parsers for the widely used search engines SEQUEST, Mascot, Spectrum Mill, X! Tandem, and OMSSA. MASPECTRAS manages the sequence databases used while searching with different modules internally. Any database available in FASTA format [28] can be uploaded to MASPECTRAS. Parsing rules are user definable and therefore easily adaptable to different types of sequence databases. When results of a search engine are imported into MASPECTRAS, the system first tries to determine whether the same accession string for the same database version exists. If that is not the case, the original sequence information is retrieved from the corresponding sequence database. Subsequently the system tries to match the sequence against the sequences already stored in the database. If an entry with the same sequence information but a different accession string is found, the new accession string is associated with the unique identifier of the already stored sequence. Otherwise a new unique identifier is created and the sequence is stored with the appropriate accession strings.

Peptide Validation

SEQUEST and Mascot provide custom probability scores. MASPECTRAS provides a probability score on its own for SEQUEST and Mascot which is based on the algorithm of PeptideProphet [22]. Data re-scoring adds a further layer, which improves the specificity of the highly sensitive SEQUEST and Mascot database searches. This procedure could be applied to other database search algorithms as well and can additionally offer a remap of the results from different database search algorithms onto one single probability scale [21]. The statistical model incorporates a linear discriminant score based on the database search scores (for SEQUEST: XCorr, dCn, Sp rank, and mass difference) as well as the tryptic termini and missed cleavages [22]. After scoring the data has to pass a user definable filter, which depends on the search programs specific score to discard the most unlikely data.

Protein Clustering

In peptide fragmentation fingerprinting (PFF) peptides are identified by search engines, which have to be mapped to proteins. A single peptide often corresponds to a group of proteins. Therefore, PFF identifies protein groups, each protein owning similar peptides. A grouped protein view represents the result more concisely and proteins with a small number of identified peptides can be recognized easier

in complex samples. The protein grouping implemented in MASPECTRAS is based on Markov clustering [29] using Basic Local Alignment Search Tool (BLAST) [30] and multiple alignments. A file in FASTA format is assembled containing all sequences to be clustered. Each sequence is then compared against each other. The all-against-all sequence similarities generated by this analysis are parsed and stored in an upper triangular matrix. This matrix represents sequence similarities as a connection graph. Nodes of the graph represent proteins, and edges represent sequence similarity that connects such proteins. A weight is assigned to each edge by taking the average pair wise $-\log_{10}$ of the BLAST E-value. These weights are transformed into probabilities associated with a transition from one protein to another within this graph. This matrix is parsed through iterative round of matrix multiplication and inflation until there is little or no net change in the matrix. The final matrix is then interpreted as the protein clustering and the number of the corresponding cluster is stored for every protein hit. The visualization of the protein grouping of a single search is performed by the integrated Jalview Alignment Editor [31]. If proteins from different searches are the same the two corresponding protein groups are combined into one protein group at the time the searches are compared.

Protein Quantification

For quantification of peptides the ASAPRatio algorithm described in [32] has been integrated and applied: To determine a peak area a single ion chromatogram is reconstructed for a given m/z range by summation of ion intensities. This chromatogram is then smoothed tenfold by repeated application of the Savitzky - Golay smooth filtering method [33]. For each isotopic peak center and width are determined. The peak width is primarily calculated by using the standard ASAPRatio algorithm and for further peak evaluation a new algorithm for recognizing peaks with saddle-points has been implemented. With this algorithm a valley (a local minimum of the smoothed signal) is recognized to be part of the peak and added to the area. The calculated peak area is determined as the average of the smoothed and the unsmoothed peak. From this value background noise is subtracted, which is estimated from the average signal amplitude of the peak's neighborhood (50 chromatogram value pairs above and below the respective peak's borders). The peak error is estimated as the difference of the smoothed and the unsmoothed peak. A calculated peak area is accepted in case the calculated peak

area is bigger than the estimated error and the peak value is at least twice the estimated background noise, otherwise the peak area is set to zero. The acceptance process is applied in automated peak area determination, only. In case of interactive peak determination this process is replaced by the operator's decision. In order to demonstrate the quantification capabilities of MASPECTRAS two samples were mixed at different ratios and quantified with MSQuant [34], PepQuan (provided with the Bioworks browser from SEQUEST), and MASPECTRAS. The results are described in "System Validation". For a detailed description of the experiment see "Experimental Procedures".

Visualization Tools

MASPECTRAS allows the storage and comparison of search results from the search engines SEQUEST, Mascot, Spectrum Mill, X! Tandem, and OMSSA matched to different sequence databases merged in a single user-definable view (Figure 2). MASPECTRAS provides customizable (clustered) protein, peptide, spectrum, and chromatogram views, as well as a view for the quantitative comparison.

The clustered protein view displays one representative for each protein cluster. In the peptide centric view the peptides with the same modifications are combined together and only the representative with the highest score is displayed. The spectrum viewer of MASPECTRAS enables manual inspection of the data by providing customizable zooming and printing features (Figure 3). The chromatogram viewer allows manual definition of the peak areas (Figure 4). The chromatograms of all charge states of the found peptide are displayed. The quantitative comparison view offers the possibility to compare peptides with two different post translational modifications (PTMs) or with one PTM and an unmodified version. The calculated peaks are displayed graphically together with a regression line.

PRIDE Export

MASPECTRAS has been designed to comply with the MIAPE requirements and provide researchers all the advantages of following standards: data can be easily exported to other file formats (Excel, Word, and plain text). MASPECTRAS features a module for the PRIDE export. The export to the

PRIDE XML format is possible directly from the protein and peptide views and the resulting file can be submitted to PRIDE.

SYSTEM VALIDATION

Analysis of large proteomics data set

To demonstrate the utility of the MASPECTRAS we used data from a large-scale study recently published by Kislinger et. al. [35]. We analyzed the data from the heart cytosol compartment which comprised 84 SEQUEST searches performed against a database obtained from the authors (see <https://maspectras.genome.tugraz.at>) containing the same amount of “decoy” proteins presented in inverted amino acid orientation. The files were imported, parsed, the data analyzed and the results exported in PRIDE format. In the study of Kislinger et. al. a protein was accepted with a minimum of two high scoring spectra with a likelihood value >95% (calculated by STATQUEST [36]), which resulted in 698 protein identifications in the cytosol compartment. Applying the same filter criteria and using the PeptideProphet algorithm implemented in MASPECTRAS resulted in 570 protein identifications (81.7%). The results of this analysis are shown in additional data file 1 and at <https://maspectras.genome.tugraz.at>.

Quantitative analysis

To evaluate the performance of the quantification tool we initiated a controlled experiment in triplicates using mixture of ICPL-labeled (Isotope Coded Protein Label) proteins (see Experimental Procedures). ICPL-labeled probes were mixed at 7 different ratios (1:1, 2:1, 5:1, 10:1, 1:2, 1:5 and 1:10). To demonstrate the capabilities of MASPECTRAS, the quantitative analysis was performed with MSQuant [34], PepQuan, and ASAPRatio as implemented in MASPECTRAS. Due to the fact that MSQuant lacks the ability to quantify samples in centroid mode, the automatic quantification of MSQuant and MASPECTRAS has been performed on profile mode data. Additionally we compared the automatic quantification of MASPECTRAS in centroid mode and observed no significant deviation (data not shown).

Since in the centroid mode the data amount is smaller (~1/8) the manual review and correction of the automatically calculated results has been conducted with centroid mode data. The reasons for the

manual correction are: (i) there are additional peaks in a chromatogram in the m/z neighborhood; (ii) the found peptides are not in the main peak but in neighboring smaller peak. A ratio between each found light and heavily labeled peptide has been calculated, and from those ratios the mean value, the standard deviation, the relative error, and a regression line has been calculated as well (with the integrated PTM quantitative comparison tool described in the “Visualization tools” section). A filter for outlier removal has been applied to the automatically calculated ratios. For the manual evaluation, these automatically removed peptides were checked manually and the misquantification due to the above mentioned reasons could be corrected. Therefore the number of manually accepted peptides could be higher than the automatically accepted ones. The performance of the quantification with ASAPRatio integrated in MASPECTRAS was superior compared to both, MSQuant and PepQuan. Furthermore, for all ratios the relative error calculated was considerably lower than the relative error obtained with MSQuant and PepQuan (see table 2 and for more detailed information additional data file 2 for a direct comparison between MSQuant, PepQuan, and MASPECTRAS).

DISCUSSION

We have developed an integrated platform for the analysis and management of proteomics LC-MS/MS data using state-of-the-art software technology. The uniqueness of the platform lies in the MIAPE compliance, PRIDE export, and the scalability of the system for computationally intensive tasks, in combination of common features for data import from common search engines, integration of peptide validation, protein grouping and quantification tools.

MIAPE compliance and PRIDE export are necessary to disseminate data and effectively analyze a proteomics experiment. As more and more researchers are adopting the standards, public repositories will not only enhance data sharing but will also enable data mining within and across experiments. Surprisingly, although standards for data representation have been widely accepted, the necessary software tools are still missing. This can be partly explained by the volume and complexity of the generated data and by the heterogeneity of the used technologies. We have therefore positioned the beginning of the analytical pipeline of MASPECTRAS at the point at which the laboratory workflows converge, i.e. analysis of the data generated by the search engines.

The capability to import and parse data from five search engines makes the platform universal and independent of the workflow performed by the proteomics research group. The system was not designed to support a specific manufacturer and can therefore be used in labs equipped with different instruments. Moreover, MASPECTRAS is the first system that provides the basis for consensus scoring between MS/MS search algorithms. It was recently suggested that the interpretation of the results from proteomics studies should be based on the analysis of the data using several search engines [21]. Importing and parsing the results from search engines and side-by-side graphical representation of the results is a prerequisite for this type of analysis and would enhance correct identification of peptides. The results of the validation of our system using large proteomics data sets further support this observation. The differences in the results of the analyses are due to the different algorithms used for the likelihood calculation. In our system PeptideProphet [22] was used whereas in

the study by Kislinger et al. [35] STATQUEST [36] was applied. We have selected PeptideProphet algorithm based on the results of the benchmarking study [21] in which PeptideProphet was ranked first with respect to the number of correctly identified peptide spectra. The study by Kapp et al. [21] showed also that the concordance between MS/MS search algorithms can vary up to 55% (335 peptides were identified by all four algorithms out of possible 608 hits). Important considerations when carrying out MS/MS database searches is not only the chosen search engine, but also the specified search parameters, the search strategy, and the chosen protein sequence database. Evaluation of the performance of the used algorithms was beyond the scope of this study. Further work need to be carried out to determine the number of independent scoring functions necessary to allow automated validation of peptide identifications. It should be noted that inclusion of additional validation algorithms in MASPECTRAS is straightforward due to the flexibility of the platform and the use of standard software technology.

The integration of peptide validation, protein grouping and quantification algorithms in conjunction with visualization tools is important for the usability and acceptability of the system. Particularly the inclusion of a quantification algorithm in the pipeline is of interest since more and more quantitative studies are initiated. We have selected the ASAPRatio algorithm for automated statistical analysis of protein abundance ratios [32] and integrated it into our platform. The results of our validation experiment showed that the performance of ASAPRatio was superior to MSQuant and PepQuan. Again, the modularity of the platform allows future integration of other quantification algorithms. Moreover, the use of three-tier software architecture in which the presentation, the calculation and the database part are separated enables not only easier maintenance but also future changes like inclusion of additional algorithms as well as distribution of the load to several servers. We made use of the flexibility of this concept and developed a module for distributing the load to a computing cluster (JClusterService, see Software Architecture). Tests with the ASAPRatio algorithm showed that the computing time decreases linearly with the number of used processors.

In summary, given the unique features and the flexibility due to the use of standard software technology, our platform represents significant advance and could be of great interest to the proteomics community.

SOFTWARE ARCHITECTURE

The application is based on a three-tier architecture, which is separated into presentation-, middle-, and database layer. Each tier can run on an individual machine without affecting the other tiers. This makes every component easily exchangeable. A relational database (MySQL, PostgreSQL or Oracle) forms the database layer. MASPECTRAS follows and extends the PEDRo database scheme [6] (additional data file 3) to suit the guidelines of PSI [4]. The business layer consists of a Java 2 Enterprise Edition (J2EE) compliant application which is deployed to the open source application server JBoss [37]. Access to the data is provided by a user-friendly web-interface using Java Servlets and Java Server Pages [38] via the Struts framework [39]. Computational or disk space intensive tasks can be distributed to a separate server or to a computing cluster by using the in-house developed JClusterService interface. This web service based programming interface uses the Simple Object Access Protocol (SOAP) [40] to transfer data for the task execution between calculation server and MASPECTRAS server. The tasks can be executed on dedicated computation nodes and therefore do not slow down the MASPECTRAS web interface. This remote process execution system is used as a backend for the protein grouping analysis, for the mass quantification and for the management of the sequence databases and their sequence retrieval during import.

The current implementation of MASPECTRAS allows the comparison of search results from SEQUEST, Mascot, Spectrum Mill, X! Tandem [25], and OMSSA [26]. The following file formats are supported: SEQUEST: ZIP-compressed file of the *.dta, *.out and SEQUEST.params files; Mascot: *.dat; Spectrum Mill: ZIP-compressed file of the results folder including all subfolders; X! Tandem: the generated *.xml; OMSSA: the generated *.xml with included spectra and search params; Raw data: XCalibur raw format (*.raw) version 1.3, mzXML [41] and mzData [42] format. The data can be imported into MASPECTRAS database asynchronously in batch mode, without interfering with the analysis of already uploaded data. The spectrum viewer applet and the diagrams are implemented with the aid of JFreeChart [43] and Cewolf [44] graphics programming frameworks. The whole system is secured by a user management system which has the ability to manage the access rights for projects and offers data sharing and multiple user access roles in a multi-user environment.

EXPERIMENTAL PROCEDURES

In order to demonstrate the capabilities of MASPECTRAS the following experiments were performed.

Materials

Proteins were purchased from Sigma as lyophilized, dry powder. Solvents (HPLC grade) and chemicals (highest available grade) were purchased from Sigma, TFA (trifluoroacetic acid) was from Pierce. The ICPL (isotope coded protein label) chemicals kit was from Serva Electrophoresis this kit contained reduction solution with TCEP (Tris (2-carboxy-ethyl) phosphine hydrochloride), cysteine blocking solution with IAA (Iodoacetamide), stop solutions I and II and the labeling reagent nicotinic acid N-hydroxysuccinimide ester as light (6 ^{12}C in the nicotinic acid) and heavy (6x ^{13}C) form as solutions. Trypsin was purchased from Sigma at proteomics grade.

ICPL labeling of proteins

Proteins bovine serum albumin [GenBank:AAA51411.1], human apotransferrin [ref:NP_001054.1] and rabbit phosphorylase b [PDB:8GPB] were dissolved with TEAB (Tetraammoniumbicarbonate) buffer (125 mM, pH 7.8) in three vials to a final concentration of 5 mg/ml each. A 40 μl aliquot was used for reduction of disulfide bonds between cysteine sidechains and blocking of free cysteines. For reduction of disulfide bonds 4 μl of reduction solution were added to the aliquot and the reaction was carried out for 35 min at 60 $^{\circ}\text{C}$. After cooling samples to room temperature, 4 μl of cysteine blocking solution were added and the samples were sat in a dark cupboard for 35 min. To remove excess of blocking reagent 4 μl of stop solution I were added and samples were put on a shaker for 20 minutes. Protein aliquots were split to two samples which contained 20 μl each. First row of samples was labeled with the ^{12}C isotope by adding 3 μl of the nicotinic acid solution which contained the light reagent. Second row was labeled with the heavy reagent and labeling reaction was carried out for 2 h and 30 min while shaking at room temperature.

Proteolytic digest of Proteins

Protein solutions were diluted using 50 mM NH_4HCO_3 solution to a final volume of 90 μl . 10 μl of a fresh prepared trypsin solution (2.5 $\mu\text{g}/\mu\text{l}$) were added and the proteolysis was carried out at 37 °C over night in an incubator. The reaction was stopped by adding 10 μl of 10% TFA. The peptide solutions were diluted with 0.1 % TFA to give 1 nM final concentration. From these stock solutions samples for MS/MS analysis which contained defined ratios of heavy and light were made up by mixing the solutions of light and heavy labeled peptides.

HPLC and mass spectrometry

To separate peptide mixtures prior to MS analysis, nanoRP-HPLC was applied on the Ultimate 2 Dual Gradient HPLC system (Dionex, buffer A: 5% ACN, 0.1% TFA, buffer B: 80% ACN, 0.1% TFA) on a PepMap separation column (Dionex, C18, 150 mm x 75 μm x 3 μm , 300 A). 500 fMol of each mixture was separated three times using the same trapping and separation column to reduce the quantification error which comes from HPLC and mass spectrometry. A gradient from 0% B to 50% B in 48 min was applied for the separation; peptides were detected at 214 and 280 nm in the UV detector. The exit of the HPLC was online coupled to the electrospray source of the LTQ mass spectrometer (Thermo Electron). Samples were analyzed in centroid mode first to test digest and labeling quality. For the quantitative analysis the LTQ was operated in enhanced profile mode for survey scans to gain higher mass accuracy. Samples were mass spectrometrically analyzed using a top one method, in which the most abundant signal of the MS survey scan was fragmented in the subsequent MS/MS event in the ion trap. Although with this method a lower number of MS/MS spectra were acquired, the increased number of MS scans leads to a better determination of the eluting peaks and therefore provides improved quantification of peptides.

Data analysis was done with the Mascot Daemon [23] (Matrix Science), BioWorks 3.2 [22] (Thermo Electron) software packages using an in house database. To demonstrate the merging of results from all of the mentioned search engines the ICPL labeled probes at an ratio of 1:1 were searched with Spectrum Mill A.03.02 (Agilent Technologies) [24], X! Tandem [25] (The Global Proteome Machine

Organization) version 2006.04.01, and OMSSA 1.1.0 [26] (NCBI) The results were uploaded to MASPECTRAS and quantified automatically.

AUTHORS' CONTRIBUTIONS

JH designed the current version of MASPECTRAS. He was responsible for the implementation of the database, the development the presentation and many parts of the business logic. GS, AS¹, TRB and EK implemented most of the parts of the analysis pipeline. GS developed the JClusterService and the services provided for MASPECTRAS. TRB integrated the PeptideProphet, AS¹ the protein clustering pipeline, and EK the peptide quantification and the chromatogram viewer. RR implemented the PRIDE data export. AS² and KM conducted the proteomics experiments. JH and AS² analyzed the biological data. KM and GGT contributed to conception and design. ZT was responsible for the overall conception and project coordination. All authors gave final approval of the version to be published.

ACKNOWLEDGEMENTS

The authors thank the staff of the protein chemistry facility at the Research Institute of Molecular Pathology Vienna, Sandra Morandell and Stefan Ascher, Biocenter Medical University Innsbruck, Manfred Kollroser, Institute of Forensic Medicine, Medical University of Graz, Gerald Rechberger, Institute of Molecular Biosciences, University of Graz, Andreas Scheucher, and Thomas Fuchs for valuable comments and contributions. We want to thank Andrew Emili and Vincent Fong from the Donnelly Centre for Cellular and Biomolecular Research (CCBR), University of Toronto for providing the data for our study. This work is supported by the Austrian Federal Ministry of Education, Science and Culture GEN-AU projects “Bioinformatics Integration Network II” (BIN) and “Austrian Proteomics Platform II” (APP). Jürgen Hartler was supported by a grant of the Austrian Academy of Sciences (OEAW).

REFERENCES

Reference List

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC et al.: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
2. Maurer M, Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, Prokesch A, Scheideler M, Trajanoski Z: **MARS: microarray analysis, retrieval, and storage system.** *BMC Bioinformatics* 2005, **6**:101.
3. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3**:SOFTWARE0003.
4. Orchard S, Hermjakob H, Apweiler R: **The proteomics standards initiative.** *Proteomics* 2003, **3**:1374-1376.
5. Orchard S, Hermjakob H, Julian RKJ, Runte K, Sherman D, Wojcik J, Zhu W, Apweiler R: **Common interchange standards for proteomics data: Public availability of tools and schema.** *Proteomics* 2004, **4**:490-491.
6. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I et al.: **A systematic approach to modeling, capturing, and disseminating proteomics experimental data.** *Nat Biotechnol* 2003, **21**:247-254.
7. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: The proteomics identifications database (vol. 5, Issue 13, pp. 3537-3545).** *Proteomics* 2005, **5**:4046.
8. Keller A, Eng J, Zhang N, Li XJ, Aebersold R: **A uniform proteomics MS/MS analysis platform utilizing open XML file formats.** *Mol Sys Biology* 2005, **4100024**:E1-E8.
9. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *J Proteome Res* 2004, **3**:1234-1242.
10. Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON: **VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins.** *J Proteome Res* 2005, **4**:2338-2347.
11. Matthiesen R, Bunkenborg J, Stensballe A, Jensen ON, Welinder KG, Bauw G: **Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V20.** *Proteomics* 2004, **4**:2583-2593.
12. Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, Hussey P, Igra M, Maclean B, Lin CW, Detter A et al.: **Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments.** *J Proteome Res* 2006, **5**:112-121.
13. Eddes JS, Kapp EA, Frecklington DF, Connolly LM, Layton MJ, Moritz RL, Simpson RJ: **CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies.** *Proteomics* 2002, **2**:1097-1103.

14. Wilke A, Ruckert C, Bartels D, Dondrup M, Goesmann A, Huser AT, Kesphohl S, Linke B, Mahne M, McHardy A et al.: **Bioinformatics support for high-throughput proteomics.** *J Biotechnol* 2003, **106**:147-156.
15. Garden P, Alm R, Hakkinen J: **PROTEIOS: an open source proteomics initiative.** *Bioinformatics* 2005, **21**:2085-2087.
16. Shadforth I, Xu W, Crowther D, Bessant C: **GAPP: a fully automated software for the confident identification of human peptides from tandem mass spectra.** *J Proteome Res* 2006, **5**:2849-2852.
17. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Edes J, Loevenich SN, Aebersold R: **The PeptideAtlas project.** *Nucleic Acids Res* 2006, **34**:D655-D658.
18. Kristensen DB, Brond JC, Nielsen PA, Andersen JR, Sorensen OT, Jorgensen V, Budin K, Matthiesen J, Venø P, Jespersen HM et al.: **Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data.** *Mol Cell Proteomics* 2004, **3**:1023-1038.
19. Shinkawa T, Taoka M, Yamauchi Y, Ichimura T, Kaji H, Takahashi N, Isobe T: **STEM: a software tool for large-scale proteomic data analyses.** *J Proteome Res* 2005, **4**:1826-1831.
20. Kohlbacher O, Reinert K, Gropl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M: **TOPP--the OpenMS proteomics pipeline.** *Bioinformatics* 2007, **23**:e191-e197.
21. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS et al.: **An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.** *Proteomics* 2005, **5**:3475-3490.
22. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.** *Anal Chem* 2002, **74**:5383-5392.
23. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**:3551-3567.
24. **Agilent Technologies** [<http://cagchem.cos.agilent.com>]
25. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra.** *Bioinformatics* 2004, **20**:1466-1467.
26. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm.** *J Proteome Res* 2004, **3**:958-964.
27. **MSDB** [<http://csc-fserve.hh.med.ic.ac.uk/msdb.html>]
28. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**:2444-2448.
29. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**:1575-1584.

30. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
31. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20**:426-427.
32. Li XJ, Zhang H, Ranish JA, Aebersold R: **Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry.** *Anal Chem* 2003, **75**:6648-6657.
33. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical recipes in C: the art of scientific computing.* Cambridge Press: New York; 1997.
34. **MSQuant** [<http://msquant.sourceforge.net/>]
35. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT et al.: **Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling.** *Cell* 2006, **125**:173-186.
36. Kislinger T, Rahman K, Radulovic D, Cox B, Rossant J, Emili A: **PRISM, a generic large scale proteomic investigation strategy for mammals.** *Mol Cell Proteomics* 2003, **2**:96-106.
37. **JBoss.com: The Professional Open Source Company** [<http://www.jboss.org>]
38. Hall M., Brown L.: *Core Servlets and Java Server Pages: Core Technologies.* A Sun Microsystems Press/Prentice Hall PTR Book; 2003.
39. **Struts** [<http://struts.apache.org/>]
40. **SOAP** [<http://www.w3.org/TR/soap/>]
41. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R et al.: **A common open representation of mass spectrometry data and its application to proteomics research.** *Nat Biotechnol* 2004, **22**:1459-1466.
42. Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK, Jr., Apweiler R: **Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005).** *Proteomics* 2005, **5**:3552-3555.
43. **JFreeChart** [<http://www.jfree.org/jfreechart/>]
44. **Cewolf** [<http://cewolf.sourceforge.net>]

FIGURE LEGENDS

Figure 1

Schematic overview of the analysis pipeline of MASPECTRAS. Search results from SEQUEST, Mascot, Spectrum Mill, X! Tandem, and OMSSA are imported and parsed. In the next steps peptides are validated using PeptideProphet [22] and the corresponding proteins clustered using ClustalW [30]. Then the peptides are quantified using the ASAPRatio algorithm [32], the results stored in the database and exported to the public repository PRIDE [7].

Figure 2

Combined view of the results from the search engines. The combined result view shows the comparison from 5 different search engines (SEQUEST, Mascot, Spectrum Mill, X! Tandem, and OMSSA) for bovine serum albumin (see experimental procedures for details). The line on the top lists the search results displayed in color. Sequence segments found only in one of the searches have the corresponding color whereas sequence segments found in multiple searches are colored red. The possible peptide modifications are shown under the protein sequence box. Three types of peptide modifications were defined: ICPL-light (K%), ICPL-heavy (K*), and oxidized methionine (MX\$). X! Tandem generates additional modifications at the N-terminus (N-term@, N-term&, and N-term"). X! Tandem does not provide the possibility to search variable modification states on one amino acid. Therefore, for the X! Tandem search a fixed modification at K(+105.02) and a variable modification (K\$+6.02) has been applied. In the last table the peptides are listed and only one representative for the peptide at this modification state is shown.

Figure 3

Spectrum viewer of MASPECTRAS. The spectrum viewer offers the selection of different ion series, the change to other peptide hits, zooming- and printing possibilities.

Figure 4

Chromatogram viewer for the quantification. The raw data is filtered with the m/z of the peptide found. The calculated chromatogram and the chromatograms of the neighborhood are displayed in the first view. The second view shows the selected chromatogram (the yellow colored one in the first view). Additional peaks can be added and stored peaks (colored red) can be removed. The manually selected peaks are displayed in green. The chromatogram viewer allows changing the m/z step-size, the number of displayed neighborhood chromatograms, and the charge state.

TABLES

	TPP [8]	GPM [9]	VEMS [10,11]	CPAS [12]	CHOMPER [13]	ProDB [14]	PROTEIOS [15]	GAPP [16]	PeptideAtlas [17]	EPIR [18]	STEM [19]	TOPP [20]	MASPECTRAS
Compliance													
MIAPE MSI compliant	-	*	-	-	-	-	✓	-	-	-	-	-	✓
MIAPE MS compliant	-	-	-	-	-	-	-	-	-	-	-	-	✓
MIAPE GE compliant	-	-	-	-	-	-	✓	-	-	-	-	-	✓
MIAPE GI compliant	-	-	-	-	-	-	✓	-	-	-	-	-	✓
MIAPE LC compliant	-	-	-	-	-	-	✓	-	-	-	-	-	✓
PRIDE export	-	-	-	-	-	-	-	plan.	-	-	-	-	✓
Data Import													
mzXML	✓	✓	conv.	✓	-	-	✓	-	✓	-	-	✓	✓
mzData	-	-	conv.	-	-	-	✓	✓	-	-	-	✓	✓
SEQUEST	pepX	-	-	pepX	✓	✓	-	-	✓	-	-	✓	✓
Mascot	pepX	-	✓	pepX	-	✓	✓	✓	-	✓	✓	✓	✓
SpectrumMill	-	-	-	-	-	-	-	-	-	-	-	-	✓
XITandem	-	✓	✓	✓	-	-	✓	✓	-	-	-	-	✓
OMSSA	-	-	-	-	-	-	-	-	-	-	-	-	✓
Data validation and visualization													
Search engine included	-	✓	✓	✓	-	?	-	✓	-	-	-	-	-
Additional validation algorithms	✓	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓	✓
Protein grouping - clustering	-	✓	✓	-	-	-	-	-	-	✓	-	-	✓
Merging of results from different search engines	-	-	?	✓	-	✓	✓	✓	-	-	-	-	✓
Customizable filtering	?	✓	-	part.	-	✓	?	SQL	-	✓	✓	?	✓
Spectrum viewer	✓	✓	?	✓	✓	?	?	-	-	-	✓	?	✓
Quantification:													
Relative peptide quantification	✓	-	✓	-	-	-	-	plan.	-	✓	✓	✓	✓
Adjustable m/z width for quantification	-	-	-	-	-	-	-	-	-	-	-	-	✓
Visualization of chromatograms	✓	-	✓	-	-	-	-	-	-	?	-	✓	✓
Visualization of surrounding chromatograms	-	-	?	-	-	-	-	-	-	-	-	✓	✓
Scalability:													
Parallel computing	-	-	-	-	-	-	-	?	-	-	-	-	✓

Table 1

Comparison of MASPECTRAS to other proteomics tools. ✓ fulfills criteria; — does not fulfill criteria; ? not enough information to answer this question; part. partially fulfills criteria; plan. planned; pepX fulfills via pepXML; n.a. not applicable; conv. after conversion; * fulfills parts of MIAPE called XIAPE; MSI Mass Spectrometry Informatics; MS Mass Spectrometry; GE Gel Electrophoresis; GI Gel Informatics; LC Liquid Chromatography; SQL filtering over SQL only;

10 heavy to 1 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	22	33	27	40
mean	4.64	8.94	8.31	9.85
stdev	4.83	4.9	2.85	2.99
relative error	104.09%	54.81%	34.30%	30.36%

ratio: heavy/light

1 heavy to 10 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	20	82	28	39
mean	7.54	7	9.77	9.29
stdev	4.94	2.51	3.96	1.92
relative error	65.52%	35.86%	40.53%	20.67%

ratio: light/heavy

5 heavy to 1 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	14	43	50	53
mean	2.94	4.27	4.16	4.67
stdev	2.3	1.69	1.56	1.12
relative error	78.23%	39.58%	37.50%	23.98%

ratio: heavy/light

1 heavy to 5 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	16	67	41	40
mean	13.36	3.74	4.25	4.84
stdev	5.18	1.36	1.15	0.93
relative error	38.77%	36.36%	27.06%	19.21%

ratio: light/heavy

2 heavy to 1 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	25	50	48	72
mean	1.048	2.17	2.07	2.03
stdev	1.15	0.7	0.71	0.54
relative error	109.73%	32.26%	34.30%	26.60%

ratio: heavy/light

1 heavy to 2 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	16	74	42	47
mean	4.24	2.07	2.11	1.94
stdev	4.97	3.04	0.63	0.3
relative error	117.22%	146.86%	29.86%	15.46%

ratio: light/heavy

1 heavy to 1 light

	MSQuant	PepQuan	MASPECTRAS	
	auto	manual	auto	manual
# peptides	15	67	98	77
mean	0.92	1.28	0.97	0.99
stdev	0.46	0.48	0.24	0.19
relative error	49.30%	37.50%	24.74%	19.10%

Table 2

Summary of quantitative analysis with MASPECTRAS, MSQuant and PepQuan. A filter for outlier removal has been applied to the automatically calculated ratios in MASPECTRAS. For the manual evaluation, these automatically removed peptides were checked manually and the misquantification due to wrong peak detection could be corrected. Therefore the amount of manually accepted peptides could be higher than the automatically accepted ones. The quantification with ASAPRatio integrated in MASPECTRAS performed superior compared to both, MSQuant and PepQuan. Furthermore, for all ratios the relative error calculated was considerably lower than the relative error obtained with MSQuant and PepQuan

ADDITIONAL DATA FILES

The following additional data files are included with the online version of this article: the evaluation of the heart cytosol data for the study by Kislinger et al. (additional data file 1); quantification results and comparison with MSquant and PepQuan (additional data file 2); and a tiff image of the database scheme of MASPECTRAS (additional data file 3). The original data files are downloadable directly at the MASPECTRAS application at <https://maspectras.genome.tugraz.at>.

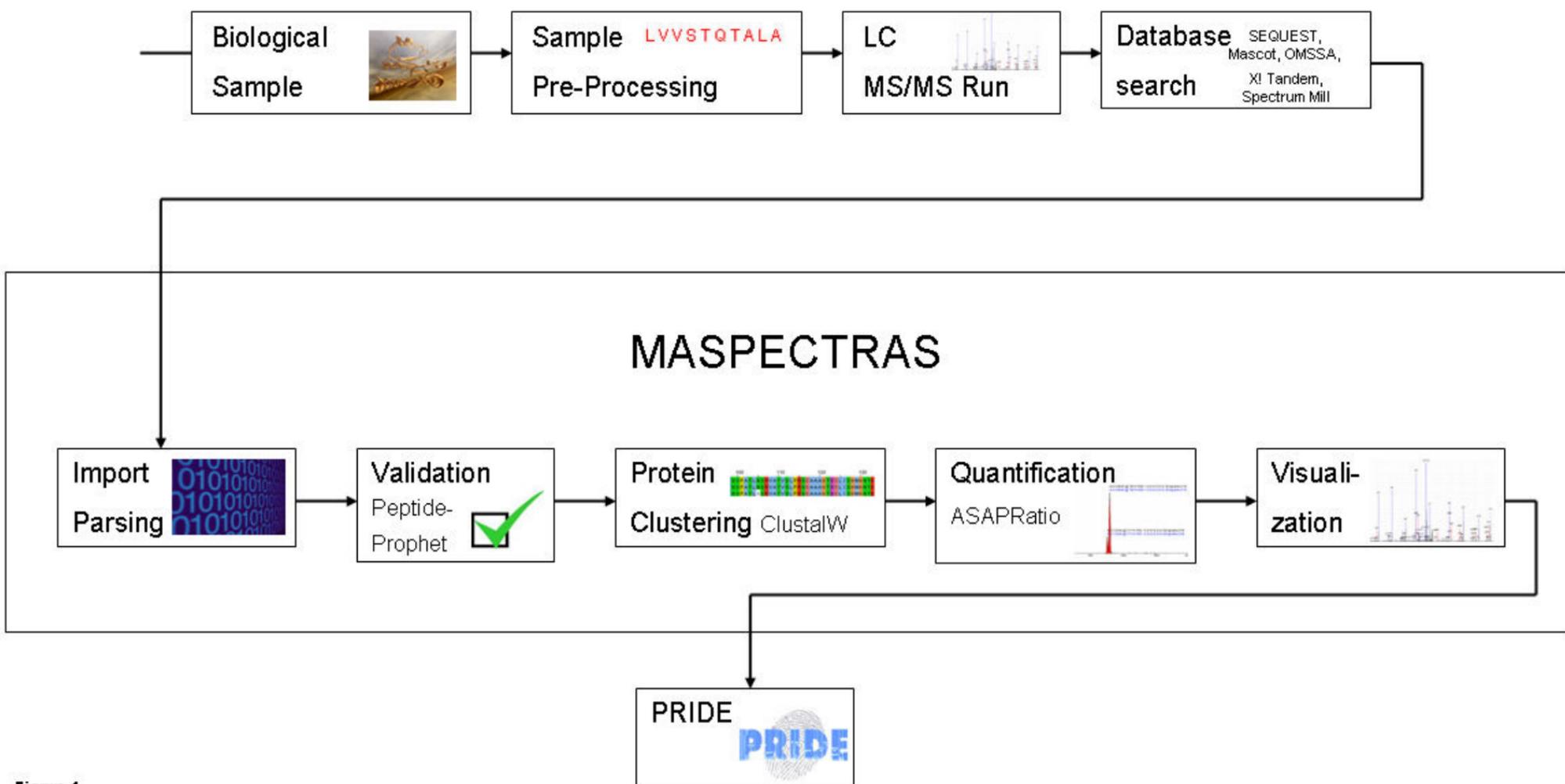


Figure 1

albumin [Bos taurus]; albumin [Bos taurus]



1 = 060606FTc2_phosphb_bsa_1hzu1IMascot 2 = 060606FTc2_phosphb_bsa_1hzu1Omssa 3 = 1hzu1ISpectrumMill
4 = 060606FTc2_phosphb_bsa_1hzu1ISequest 5 = 060606FTc2_phosphb_bsa_1hzu1IXTandem

sequence segments found in multiple searches are colored in red

Sequence

```
MKWVTFISLLLLFSSAYS RGVFRR DTHKSEIAHRFKDLGEEHF KGLVLI AF S QYLQQ C P F D E H V K L V N E  
L T E F A K T C V A D E S H A G C E K S L H T L F G D E L C K V A S L R E T Y G D M A D C C E K Q E P E R N E C F L S H K D D S P D L P K L  
K P D P N T L C D E F K A D E K K F W G K Y L Y E I A R R H P Y F Y A P E L L Y Y A N K Y N G V F Q E C C Q A E D K G A C L L P K I E T M R  
E K V L A S S A R Q R L R C A S I Q K F G E R A L K A W S V A R L S Q K F P K A E F V E V T K L V T D L T K V H K E C C H G D L L E C A D D  
R A D L A K Y I C D N Q D T I S S K L K E C C D K P L L E K S H C I A E V E K D A I P E N L P L T A D F A E D K D V C K N Y Q E A K D A F  
L G S F L Y E Y S R R H P E Y A V S V L L R L A K E Y E A T L E E C C A K D D P H A C Y S T V F D K L K H L V D E P Q N L I K Q N C D Q F E  
K L G E Y G F Q N A L I V R Y T R K V P Q V S T P T L V E V S R S L G K V G T R C C T K P E S E R M P C T E D Y L S L I L N R L C V L H E K  
T P V S E K V T K C C T E S L V N R R P C F S A L T P D E T Y V P K A F D E K L F T F H A D I C T L P D T E K Q I K K Q T A L V E L L K H K  
P K A T E E Q L K T V M E N F V A F V D K C C A A D D K E A C F A V E G P K L V V S T Q T A L A
```

All found in Red

fixed modifications

060606FTc2_phosphb_bsa_1hzu1IMascot:

Carbamidomethyl (C)

060606FTc2_phosphb_bsa_1hzu1Omssa:

carbamidomethyl C(C)

1hzu1ISpectrumMill:

Carbamidomethylation(C)

060606FTc2_phosphb_bsa_1hzu1ISequest:

(C)

060606FTc2_phosphb_bsa_1hzu1IXTandem:

(K),(C)

K*: 111.04 N-term@: 42.01 K%: 105.02 MX\$: 15.99 K\$: 6.02 N-term&: -18.02 N-term": -17.03

Peptidehits per page: 15 [25] 50 100

72 Peptidehits found

Page 1 of 3 | Next >>

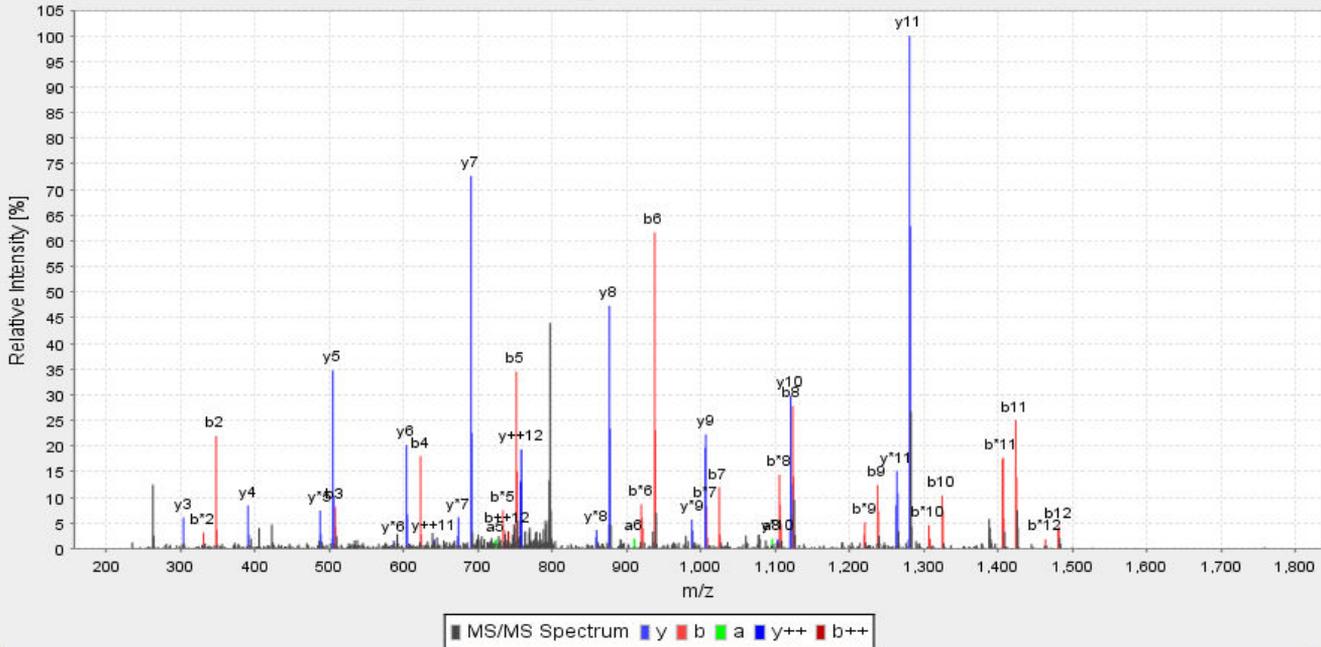
go to page go

	Search	Score	Sequence	
<input checked="" type="checkbox"/>	1 2 3	2931.6749108729373	.ALK%AWSVAR.	
<input checked="" type="checkbox"/>	1 2 3 5	2931.5490195998573	.HPEYAVSVLLR.	
<input checked="" type="checkbox"/>	1 2 3 5	2929.586073148418	.M\$PCTEDYLSLILNR.	

LK@CDEWSVNSVVGK



AS_Proteinmix_1lizu1he_A_c1.2554.2.dta



	a	b	b*	b0	b++	y	y*	y0	y++		
1	86.09	114.09	97.06	96.08	57.54	L				13	
2	319.21	347.2	330.18	329.19	174.1	K	1513.67	1496.64	1495.66	757.34	12
3	479.24	507.23	490.21	489.22	254.12	C	1280.55	1263.53	1262.54	640.78	11
4	594.27	622.26	605.23	604.25	311.63	D	1120.52	1103.5	1102.51	560.76	10
5	723.31	751.3	734.28	733.29	376.15	E	1005.5	988.47	987.48	503.25	9
6	909.39	937.38	920.36	919.37	469.19	W	876.45	859.43	858.44	438.73	8
7	996.42	1024.41	1007.39	1006.4	512.71	S	690.37	673.35	672.36	345.69	7
8	1095.49	1123.48	1106.46	1105.47	562.24	V	603.34	586.32	585.33	302.17	6
9	1209.53	1237.53	1220.5	1219.52	619.26	N	504.27	487.25	486.26	252.64	5
10	1296.56	1324.56	1307.53	1306.55	662.78	S	390.23	373.2	372.22	195.62	4
11	1395.63	1423.63	1406.6	1405.62	712.31	V	303.2	286.17	285.19	152.1	3
12	1452.65	1480.65	1463.62	1462.64	740.83	G	204.13	187.1	186.12	102.57	2
13						K	147.11	130.08	129.1	74.06	1

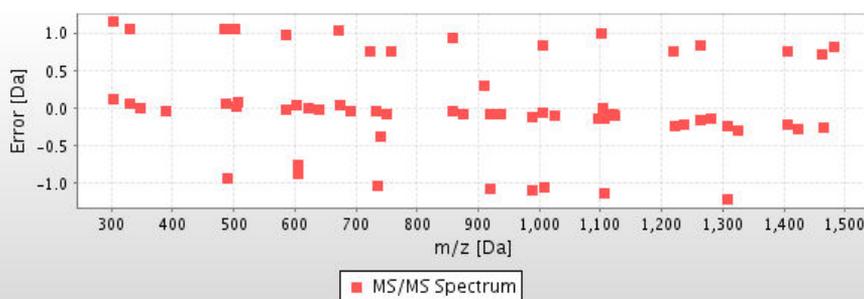


Figure 3

.ALK*AWSVAR.

V0.95

Backmost mz = 554.71 d

Frontmost mz = 559.51 d

Total Area = 1.020e+07

+ Gain

- Gain

Upper mz Span:

2

mz Step:

1.20

Lower Mz Span:

2

Process Data

Store Data

- mz

Charge:

2

+ mz

Raw

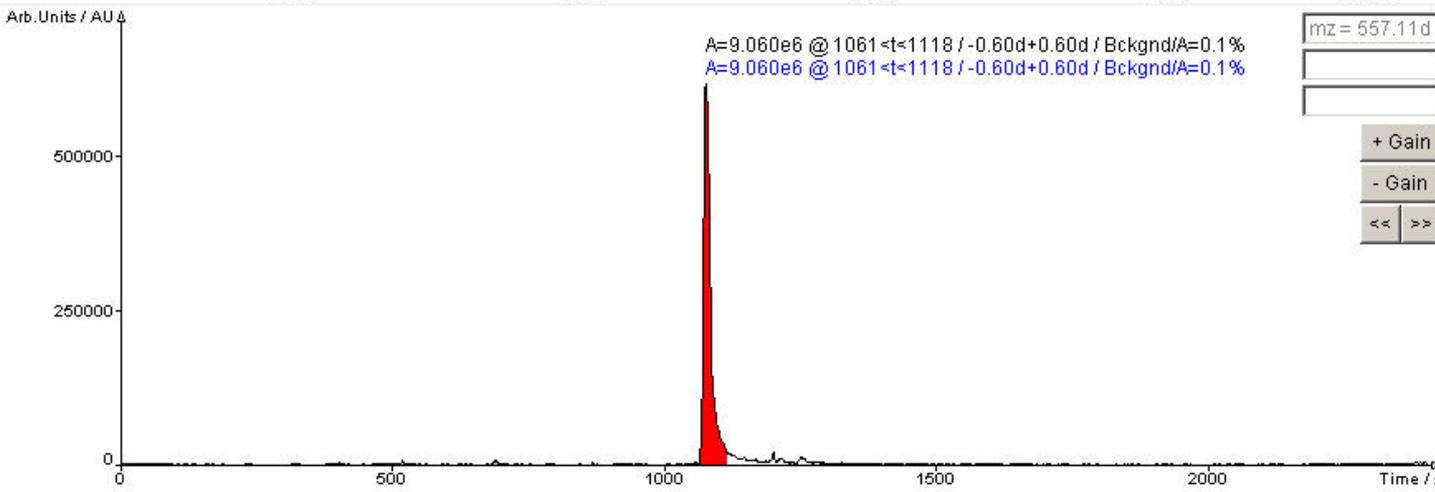
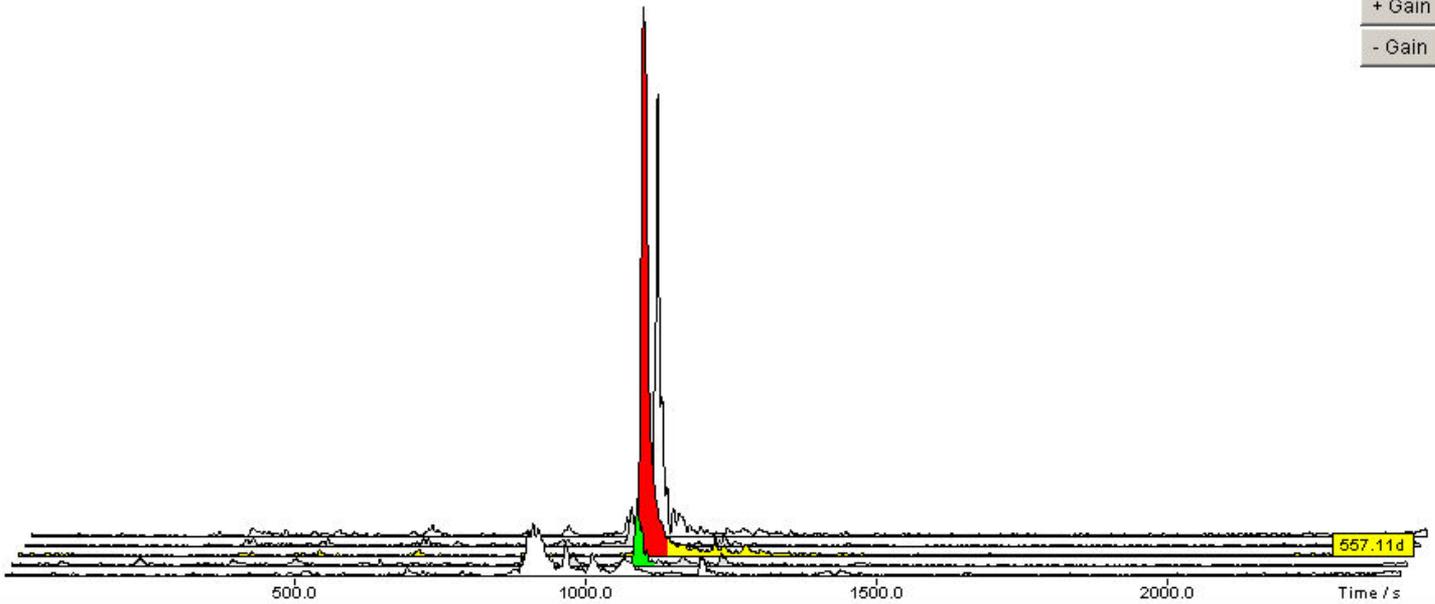
Smooth

t[min]:

t[max]:

Zoom in

Zoom all



Status: Done.

Figure 4

Additional files provided with this submission:

Additional file 1: adddata1.txt, 27K

<http://www.biomedcentral.com/imedia/9498783881405709/supp1.txt>

Additional file 2: adddata2.zip, 357K

<http://www.biomedcentral.com/imedia/7722822914057092/supp2.zip>

Additional file 3: adddata3.tif, 1010K

<http://www.biomedcentral.com/imedia/1479938682140570/supp3.tif>

Software

Open Access

MARS: Microarray analysis, retrieval, and storage system

Michael Maurer[†], Robert Molidor[†], Alexander Sturn[†], Juergen Hartler, Hubert Hackl, Gernot Stocker, Andreas Prokesch, Marcel Scheideler and Zlatko Trajanoski*

Address: Institute for Genomics and Bioinformatics and Christian Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria

Email: Michael Maurer - michael.maurer@gmail.com; Robert Molidor - robert.molidor@tugraz.at; Alexander Sturn - alexander.sturn@tugraz.at; Juergen Hartler - juergen.hartler@tugraz.at; Hubert Hackl - hubert.hackl@tugraz.at; Gernot Stocker - gernot.stocker@tugraz.at; Andreas Prokesch - andreas.prokesch@tugraz.at; Marcel Scheideler - marcel.scheideler@tugraz.at; Zlatko Trajanoski* - zlatko.trajanoski@tugraz.at

* Corresponding author †Equal contributors

Published: 18 April 2005

Received: 03 February 2005

BMC Bioinformatics 2005, 6:101 doi:10.1186/1471-2105-6-101

Accepted: 18 April 2005

This article is available from: <http://www.biomedcentral.com/1471-2105/6/101>

© 2005 Maurer et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Microarray analysis has become a widely used technique for the study of gene-expression patterns on a genomic scale. As more and more laboratories are adopting microarray technology, there is a need for powerful and easy to use microarray databases facilitating array fabrication, labeling, hybridization, and data analysis. The wealth of data generated by this high throughput approach renders adequate database and analysis tools crucial for the pursuit of insights into the transcriptomic behavior of cells.

Results: MARS (Microarray Analysis and Retrieval System) provides a comprehensive MIAME supportive suite for storing, retrieving, and analyzing multi color microarray data. The system comprises a laboratory information management system (LIMS), a quality control management, as well as a sophisticated user management system. MARS is fully integrated into an analytical pipeline of microarray image analysis, normalization, gene expression clustering, and mapping of gene expression data onto biological pathways. The incorporation of ontologies and the use of MAGE-ML enables an export of studies stored in MARS to public repositories and other databases accepting these documents.

Conclusion: We have developed an integrated system tailored to serve the specific needs of microarray based research projects using a unique fusion of Web based and standalone applications connected to the latest J2EE application server technology. The presented system is freely available for academic and non-profit institutions. More information can be found at <http://genome.tugraz.at>.

Background

Microarray analysis has become a widely used technique for the study of gene-expression patterns on a genomic scale [1,2]. Oligonucleotide and cDNA arrays have been utilized to study mRNA [3] and protein levels [4], to deci-

pher protein-DNA interactions [5], to analyze the DNA copy number [6], to detect methylated sequences [7], and to analyze gene phenotypes in living mammalian cells [8]. Microarrays represent a very complex, multi step technique involving array fabrication, labeling, hybridization,

and data analysis. Currently, most laboratories are using either one labeled sample (Affymetrix microarrays) or two labeled samples (cDNA microarrays) for hybridizations, but several applications have been established where three color microarrays are used [9,10]. State-of-the-art microarrays can have from several hundred up to tens of thousands of elements annotated by dozens of parameters. Information on details of the bench work, typically kept in lab notebooks or scattered files, as well as information regarding spotting, reliable tracking of the spotted molecules, scanning, and image quantification settings, is important for the computational analysis and reproducibility of experiments. Every step generates a wealth of data spanning tens of megabytes and in each of them errors may occur or protocols might need optimization to improve results. Moreover, all these information must be archived according to accepted scientific standards, which allow scientists to share common information and to make valid comparisons among experiments. For this reason the Microarray Gene Expression Data Society (MGED) [11] is focusing on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards. MGED is heavily promoting the sharing of high quality, well annotated data within the life sciences community. Their initiatives – MIAME (Minimum Information About a Microarray Experiment) [12], MGED Ontology [13], and MAGE-ML (MicroArray Gene Expression Markup Language) [14] – maximize the value of microarray data by permitting greater opportunities for sharing information within scientific groups and thus for discovery. These will ultimately affect the description, analysis, and management of all high throughput biological data.

The 'list of genes' resulting from microarray analysis is not the end of a microarray experiment. The major challenge is to assign biological function and to generate new hypotheses. The simplest way to find genes of potential biological interest is to search the normalized data for the highly expressed ones. Additionally, identifying patterns of gene expression and grouping genes into expression classes can provide greater insight into their biological relevance. For this purpose several supervised or unsupervised clustering algorithms like support vector machines (SVM), hierarchical clustering, k-means, self organizing map (SOM), or principal component analysis (PCA) are in use. The annotation of genes or gene clusters can be achieved by mapping them to the Gene Ontology (GO) [15] in order to provide insights into relevant molecular functions, biological processes, and cellular components [16]. Another way to identify genes of biological interest is to map the normalized data or gene expression clusters [17] to known metabolic pathways as provided e.g. by KEGG [18] or BioCarta [19].

Several academic as well as commercial systems are available that address at least some of the needs such as laboratory information management systems (LIMS) [20], microarray databases [21-24] and repositories, normalization, clustering, pathway or GO mapping tools or expression analysis platforms [25]. However, freely available systems which integrate all the aspects mentioned above are rare and may lack important issues like usability, scalability, or standardized interfaces. Furthermore, for such integrated systems it is desirable to use a uniform and state-of-the-art software architecture in order to enhance setup, maintenance and further development.

We have therefore developed a Microarray Analysis and Retrieval System (MARS) using latest Java 2 Platform, Enterprise Edition (J2EE) software technology. MARS provides modules mandatory for microarray databases:

- a laboratory information management system (LIMS) to keep track of information that accrues during the microarray production and biomaterial manipulation
- MAGE-ML export of data for depositing to public repositories e.g. ArrayExpress [26], GEO [27]

For these components already existing projects [21,23,26] have been evaluated. Their advantages as well as disadvantages have been taken into account for the design of MARS. Widely used concepts have been taken into consideration and accepted standard libraries like MAGE-STK [11] have been used whenever possible. Additionally, we extended this solid foundation and added novel features which can be highlighted as distinct advantages of the MARS system.

- a quality management application storing necessary quality control parameters indispensable for high-quality microarray data
- Web services to connect several well established tools such as normalization, clustering and pathway annotation applications
- applications for microarray normalization, gene expression clustering, and pathway exploration that are tightly integrated into the microarray analysis pipeline
- a novel, comprehensive, and Web based user management system to administrate institutes, groups, users, and their corresponding access rights

Implementation

Software architecture

MARS is based on a three tier architecture (Figure 1) using the Java 2 Platform, Enterprise Edition (J2EE), which

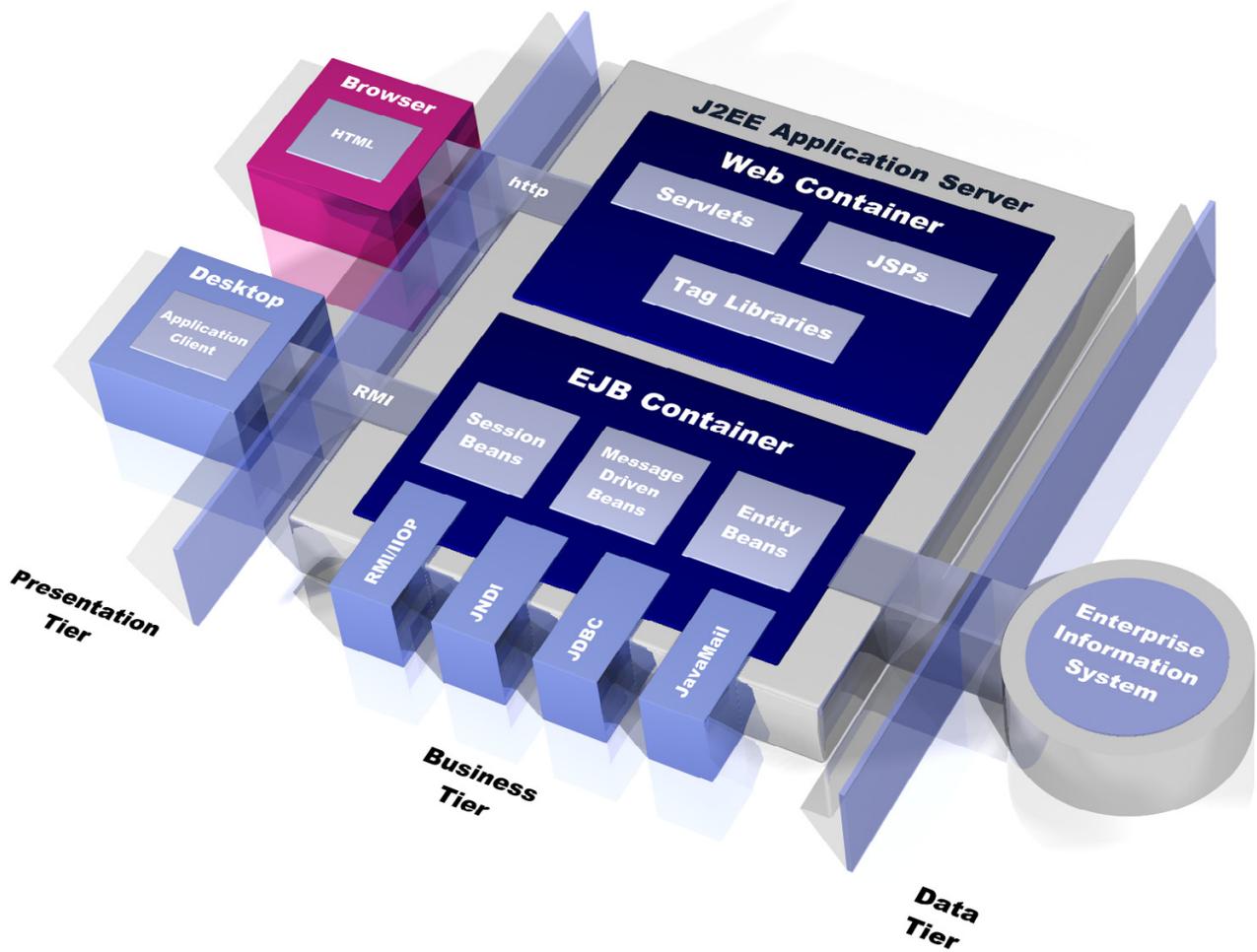


Figure 1
Three tier Java 2 Enterprise Edition software architecture. The J2EE platform simplifies the development of enterprise applications by providing standardized modular components like EJBs, JSP and Servlets. Furthermore it is providing a complete set of services to those components.

defines a standard for developing multi tier enterprise applications. The J2EE platform simplifies the development of enterprise applications using on standardized, modular components like Enterprise JavaBeans (EJB), Java Servlets, Java Server Pages (JSP), and XML technology.

A relational database (Oracle or PostgreSQL) builds the data- or Enterprise Information System tier. In the middle tier the J2EE compliant application server JBoss [28] is situated. It manages the access to the relational database as well as the interaction with the data. The Web server in conjunction with a servlet-container is responsible for the presentation tier. All the servlets and JSPs are executed to enable input and output of an application and to manage

the applications workflow logic. An advantage of a multi tier architecture is that different tiers can be deployed to different servers, enabling load distribution as well as scalability.

Systems

The database schema, the business logic, and the Web interface can be subdivided into five major groups:

1. Microarray production

To address the needs of many laboratories which produce their own microarrays, MARS includes a generic array production LIMS. It manages data regarding the substances (clones) and their localization in microtiter plates, the array design spotted on the support, as well as single

arrays and array batches. The flexible and generic database design facilitates mapping of the steadily changing laboratory workflow. Additionally, each plate can be assigned to a library, which designates the organism and contains details about the cloning vector, forward and reverse primer and standard molecule annotations including gene name, accession number, UniGene number, and sequence. Substances stored in microtiter plates may undergo certain manipulations such as PCR amplification. Therefore a PCR amplification event can be assigned to a plasmid plate in order to generate a PCR plate in the database.

After entering the information necessary for spotting, a file is generated and prepared for download. This file is used by the spotting robot software to generate an array design file. After the spotting run has been completed, the array design file has to be uploaded into MARS. For each spotting run an array batch has to be created in MARS, and all slides spotted by this spotting run have to be assigned to this array batch. Additionally, important parameters regarding the spotting run such as temperature, duration, or humidity can be assigned to this array batch. Barcode tracking is employed for plates as well as for arrays to reduce possible input errors. Laboratories using commercial arrays have to upload the array design instead and define an array batch afterwards.

2. Sample preparation

Samples can be annotated in a user-customizable manner. MARS allows the annotation of biological descriptions such as the source and characteristics of a sample (e.g. tissue and disease), any genetic and chemical manipulation and stimulation. Performing such annotations in free text fields leads to large undefined vocabularies and makes them difficult to query. Thus, three different annotation types are provided: 1) enumeration enabling the usage of defined vocabularies or ontologies, 2) numbers to allow scoring and counting and 3) free text. Annotated samples will be linked to an extract, enabling a lab worker to annotate the extraction method, protocol, concentration, purity, and quantity. The labeled extract stores information on used extract quantity, the label and the labeling protocol.

3. Hybridization and raw data management

The hybridization page archives parameters regarding the hybridization tool and method and is linked to the used labeled extracts. In contrast to several other microarray databases MARS can handle any number of labeled extracts and thus allows the storage of multi color experiments. Resulting images from hybridized scanned slides can be uploaded to MARS and added to a hybridization record. It is noteworthy that a hybridization can have several image sets associated with images of different scanner

settings. After analyzing the images several different raw datasets analyzed with different program settings can be uploaded and added to the appropriate image set.

4. Experiment annotation

A set of hybridizations forms an experiment. To store the experimental design these hybridizations can be divided into classes, paired, and flagged as a dyeswap hybridization. Additionally, an experiment can be annotated using MGED Ontology definitions (Figure 2) to specify the perturbational, methodological, and epidemiological design, as well as the biological properties. Transformed datasets can be added to classes and their corresponding raw dataset.

5. Quality management

To ensure high quality data and to allow the detection of possible sources of errors, a powerful quality management system has been integrated into MARS. This system is based on standard quality control procedures conducted during microarray production as well as during sample preparation, extraction and hybridization. In order to control the quality of PCR and purified PCR products generated during probe production, authorized users can upload gel images and analyze the bands according to a predefined schema (Figure 3). Based on this schema, PCR products can be identified later as a source of bad or missing spots on a slide. Quality annotation can be viewed by any user.

Slides can be scanned after fixation and/or after staining and parameters like spot walking or the number of missing spots are used to determine slide quality. In addition to array production quality controls, it is also necessary to check the quality of samples and its extracts. Data gained from an Agilent Bioanalyzer or gel images can be uploaded and analyzed either automatically (Bioanalyzer file) or manually (gel images) (Figure 4). Labeled extracts can be measured with a spectrophotometer to assess the efficiency of dye incorporation. Results of these measurements can be entered into MARS and the corresponding efficiency is calculated automatically. Finally, the quality of a hybridized slide is analyzed by extracting and displaying several statistical parameters from the raw data result file and by examining positive and negative controls printed onto a slide.

Data interfaces

One of the most important parts for the acceptance of a database is the data import interface. To allow the import of generic file formats, we have implemented a user definable parser that allows to read any tab delimited text file. The user has to define a file format where file columns are assigned to appropriate database fields. MARS allows to

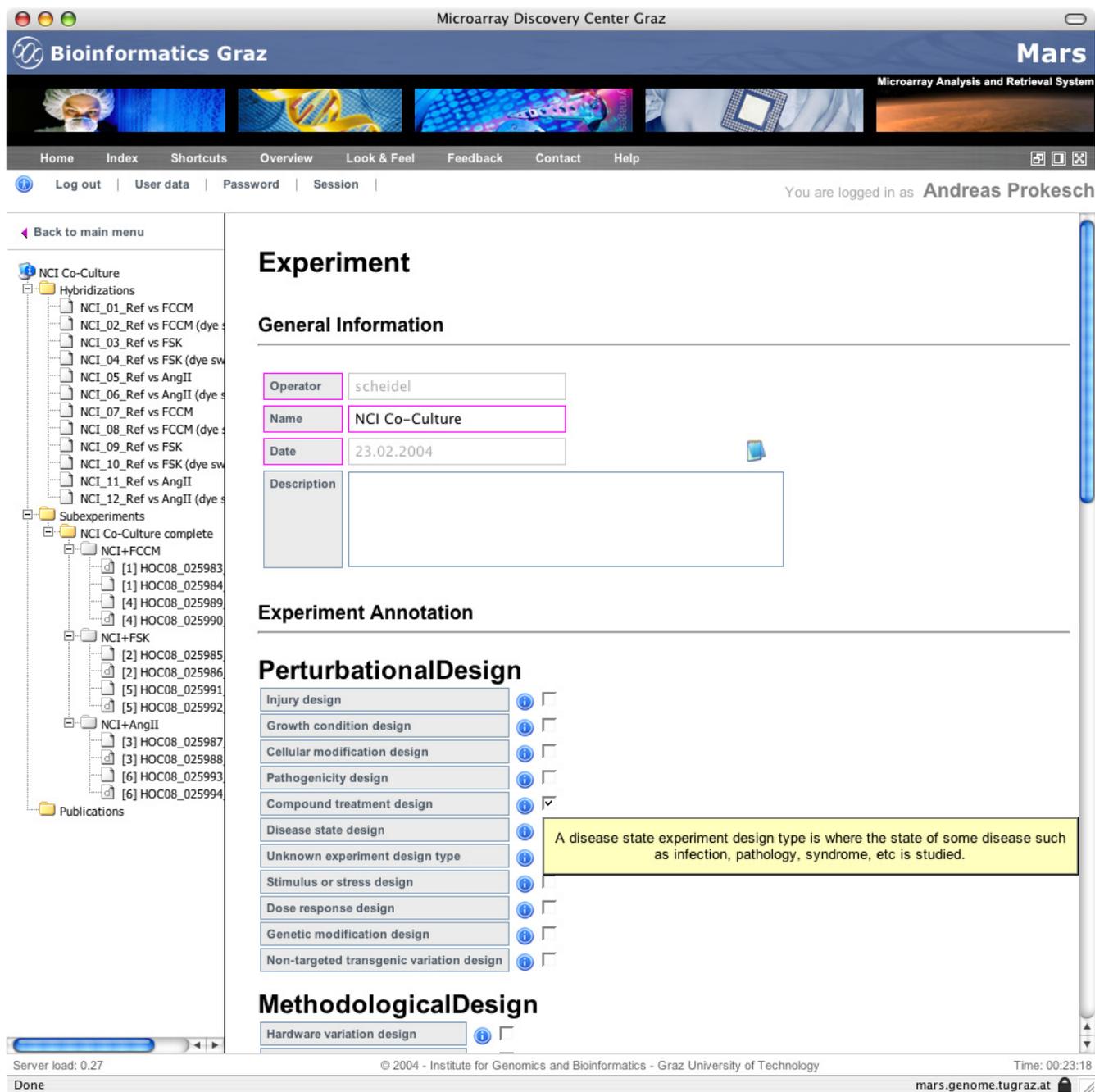


Figure 2 Experiment annotation. Web interface to define microarray experiments according to the MGED Ontology.

define file formats for importing plates, raw datasets, transformed datasets, and array designs.

Any file that has to be imported, linked, or used has to be uploaded to MARS at first. Afterwards these data can be analyzed by the users at their office desk without having to use another central storage system. Uploaded files are

stored on the servers file system where MARS has been installed. Additionally, links to these files are maintained in the relational database to prevent the deletion of already imported, linked, or used files.

The implementation of other Web based applications and more important, the usage and correct linkage of their

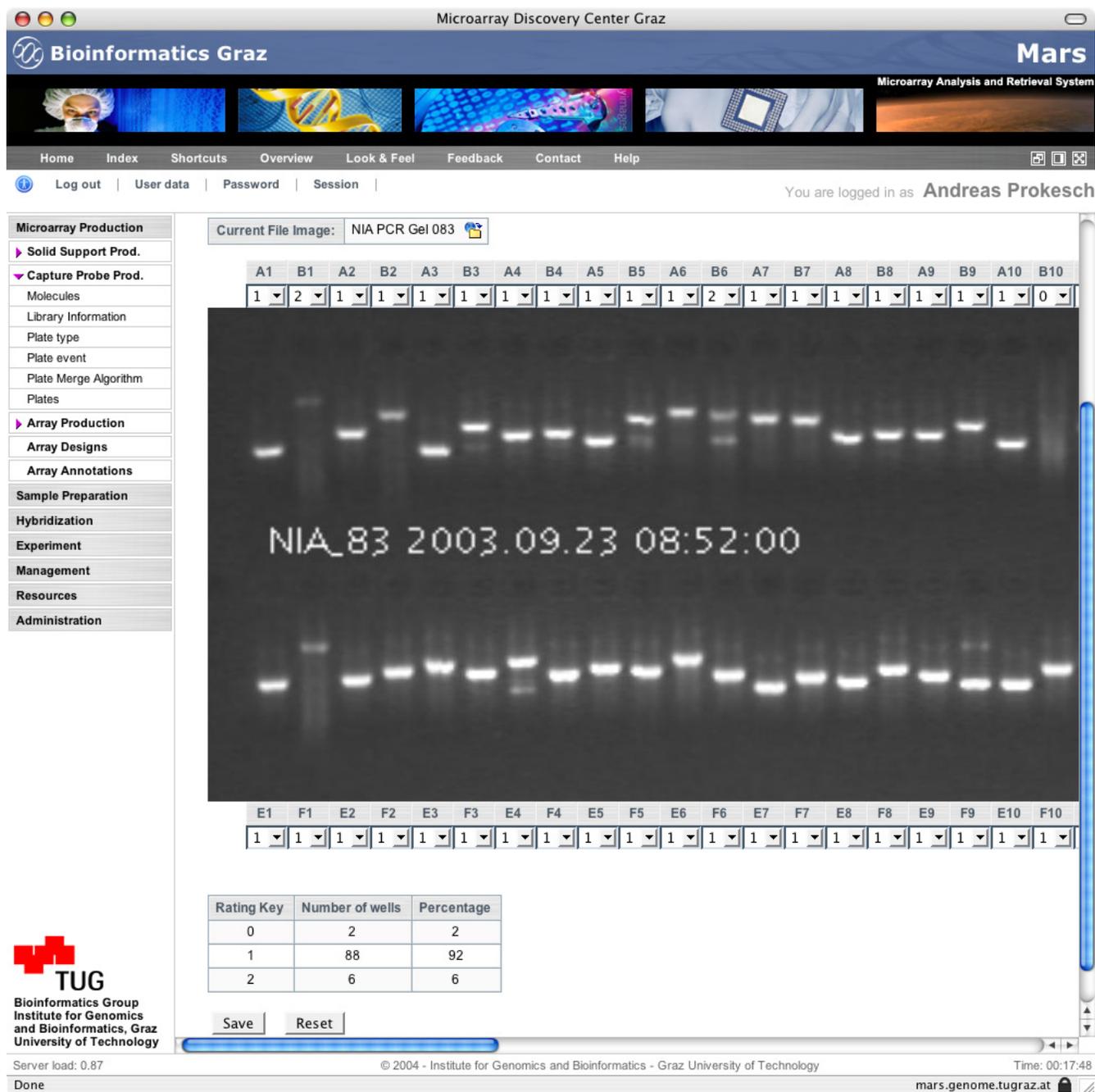


Figure 3
Quality control. A gel image from PCR products can be scored and associated to a plate.

stored data have been addressed by an External Application Connector Interface. Additional applications like supplementary quality checks can be added without any additional coding in MARS. The MARS user interface is dynamically displaying links to all former registered applications.

The Microarray Gene Expression Markup Language (MAGE-ML) has emerged as a language to describe and exchange information about microarray based experiments [29]. MAGE-ML is based on XML (eXtensible Markup Language) and can describe microarray designs, microarray manufacturing information, microarray

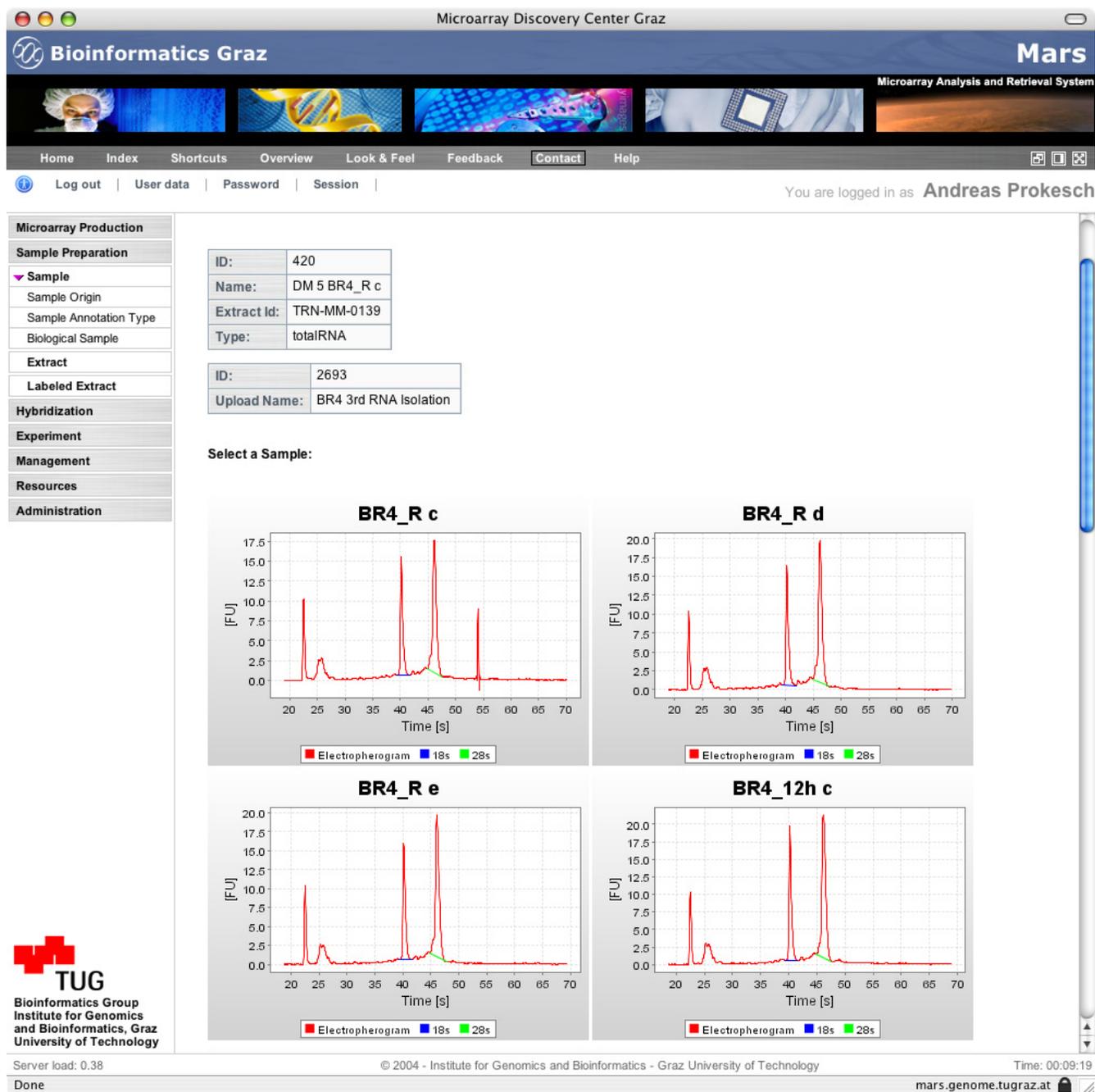


Figure 4
Quality control. Bioanalyzer analysis to check the RNA quality for a given RNA extract.

experiment setup and execution information, gene expression data, and data analysis results. By using the Java MAGE-STK (Mage Software Toolkit) [11] MARS is able to export samples, extracts, labeled extracts, array designs, raw datasets, or whole experiments including several hybridizations.

Web service

In order to grant users access to MARS with software they are familiar with (e.g. BioConductor [30] or Matlab [31]), MARS provides a well defined Simple Object Access Protocol (SOAP) interface. SOAP is an XML-based communication protocol and encoding format for inter-application

communication. After minor software adaptations these interfaces allow to authenticate against MARS, to browse own and shared datasets, to download raw data, to filter the data, and to insert transformed datasets into MARS. To take advantage of the SOAP Web service we provide a Java library called MARSExplorer, that allows software developers to extend their programs with data access functionality to MARS. Additionally, if no firewall is located between the client software and MARS, the MARS API (Application Programming Interface) can be used to access public accessible methods via the RMI (Remote Method Invocation) interface.

Access control

To avoid unauthorized database access in a multi user environment the control of user access is a crucial criterion for the acceptance of any database managing functional genomic data. Furthermore, the definition of several fine grained user access levels that allow to visualize, edit or delete data (e.g. expression and sample data, protocols) based on the user rights is mandatory. Therefore we have developed an extensible and easy to use authentication and authorization system (AAS) which rests upon the same technology as MARS. In addition to its Web based management interface, the AAS provides software libraries that enable existing and new applications the integration of highly sophisticated authentication and authorization mechanisms. Moreover, the AAS provides single-sign-on to all its connected Web based applications. Since this AAS can also be used in various projects or institutions relying upon freely available software, MySQL has been chosen as database management system. If desired, this AAS can also manage Windows and Unix accounts using SAMBA [32] and LDAP (Lightweight Directory Access Protocol) [33]. For instance, at the Institute for Genomics and Bioinformatics all Web based applications and user accounts are administrated by one single instance of the AAS.

Results

Database

All MARS user interfaces are providing a consistent look and feel and are very intuitive to use. In general, the Web based user interface can be divided into two types of user interaction pages: The first one is an input form, where a user can record required and optional data according to the MIAME standard. Required fields are marked in magenta and are validated for correct input. The second allows to list all stored records. To keep the information on a page simple, a user can hide unnecessary datafields. Furthermore it is possible to query for specific records (Figure 5) using the MARS report query tool. Because all Web pages are linked together, MARS permits to follow all conducted steps from the transformed data back to the corresponding well in a microtiter plate and to visualize

the results of quality controls. The description of an experiment including hybridizations and their raw datasets is typically the starting point for further analysis.

Analytical pipeline

The usability of MARS and the functionality of the provided interfaces and APIs (Figure 6) are revealed by the integration of MARS into an analytical pipeline of microarray analysis, beginning with image analysis, normalization, gene expression clustering, and finally mapping of gene expression data onto biological pathways.

After entering all required information into MARS, the first step is to normalize the raw data gathered from the image analysis software in order to remove systematic and random errors inherent in the data. ArrayNorm [34], an application for visualization, normalization and analysis of two-color microarray data facilitates these essential steps. Raw data including the definition of experiment classes (biological conditions) and pairs (replicated or dye swapped slides) from whole experiments can be loaded from MARS into ArrayNorm. After visualization and applying different normalization methods like linear regression, LOWESS, or self-normalization, the transformed intensities can be written back to MARS, including the history of the applied methods. The next step in the analytical pipeline is usually gene expression cluster analysis to extract the fundamental patterns inherent in the data and to organize genes with similar expression patterns into biological relevant clusters. Normalized gene expression data can be loaded into Genesis [35]. Genesis allows to cluster the dataset using various similarity distance measurements and different clustering algorithms like hierarchical clustering, k-means, self-organizing maps, principal component analysis, correspondence analysis, and support vector machines. Moreover it is possible to perform one-way ANOVA to identify differentially expressed genes and to incorporate the Gene Ontology (GO) to map gene expression clusters to GO terms. Results can be written back into MARS.

Finally, the Pathway Editor [36] provides the opportunity to access MARS and to map data either from whole experiments or from gene expression clusters to specified pathways in order to get an overview of gene expression changes and their influencing factors. All aforementioned applications have integrated MARSExplorer to connect to MARS and to query, up- and download datasets.

Discussion

The database design, state-of-the-art software technology, well designed user interface, and its application interfaces make MARS a powerful tool for storing, retrieving, and analyzing multi color microarray data. The fusion of Web based and standalone applications provides researchers

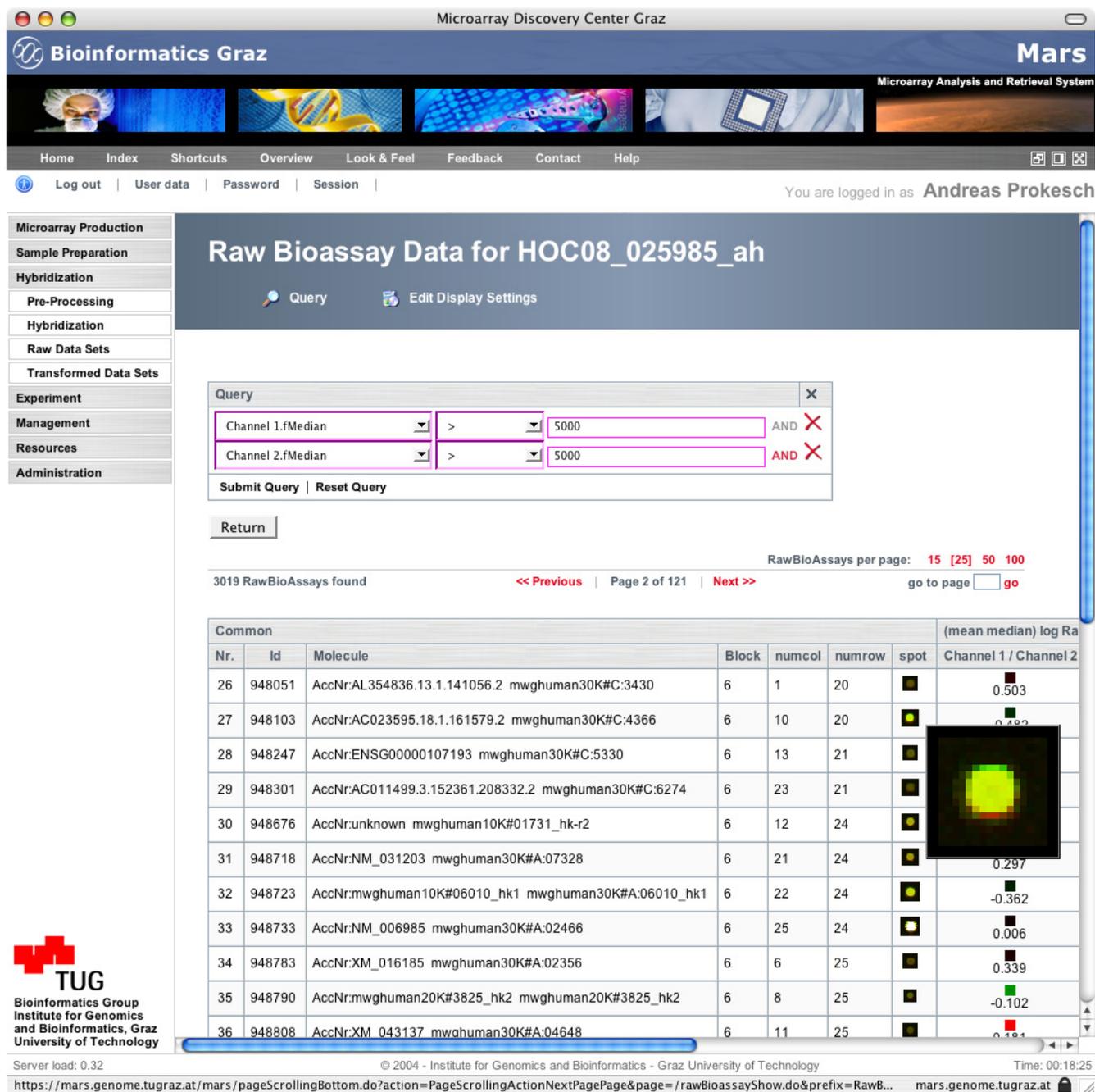


Figure 5 Typical MARS interface listing stored records. It allows to query for specific records using the user friendly query tool.

with a unique set of computational tools for genomic and transcriptomic data.

The main strengths of MARS are:

1. Data interfaces

Fundamental for the acceptance of a database are the data interfaces. In principle two types of data interfaces for human computer interactions can be distinguished. Standalone applications allow better program-user interactions while having the drawback that several or even very

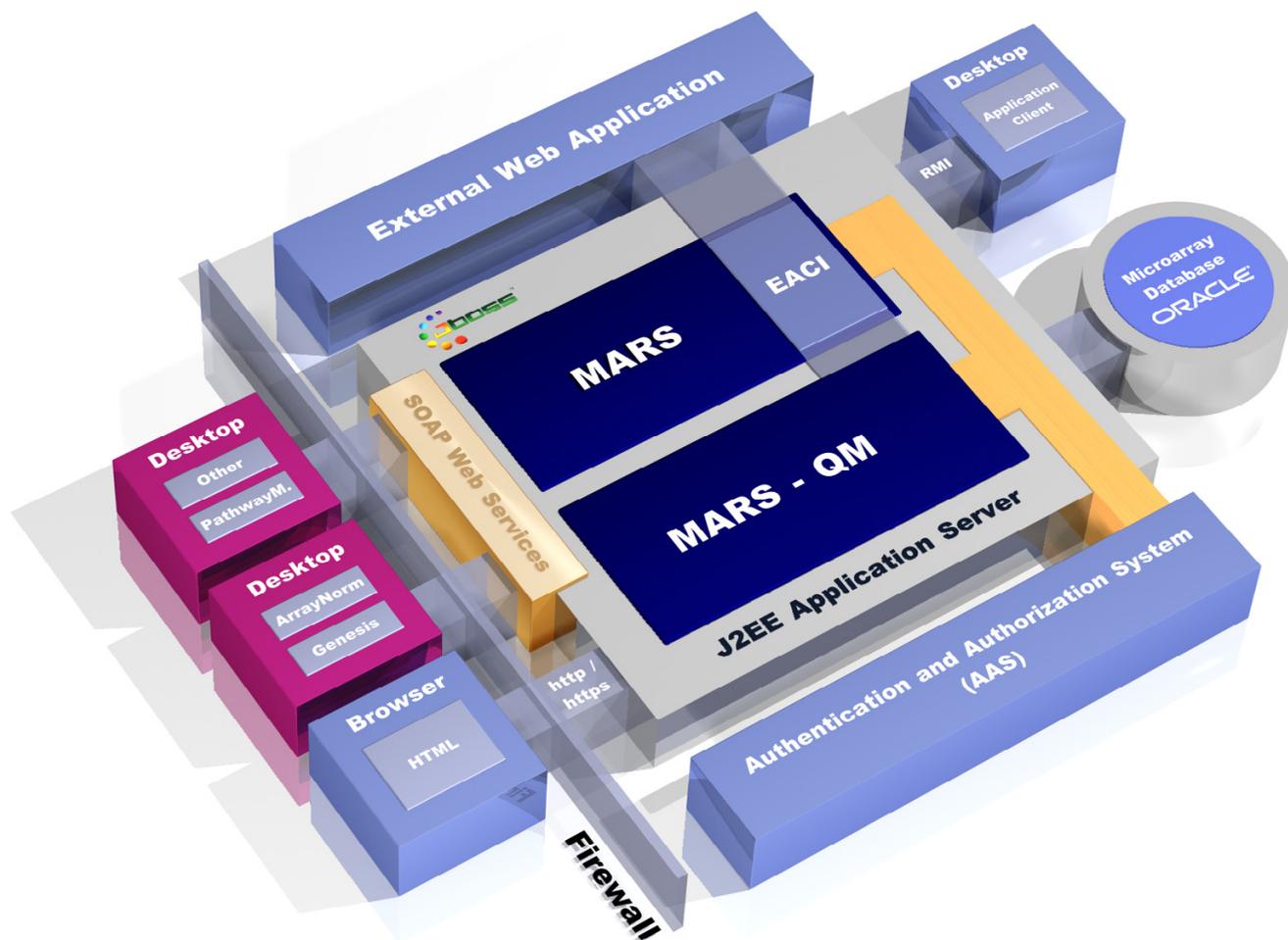


Figure 6
MARS system interactions. MARS and MARS-QM are deployed in a J2EE compliant application server. Interaction is possible either with a standard Web browser or an application supporting the SOAP or RMI protocol. The External Application Connector Interface (EACI) facilitates to connect to data from additional Web applications. SOAP and http/https enable MARS access also through firewalls.

old versions are in use. On the other hand Web based applications can be easily used on every computer without any installation effort and they provide the same and newest version to all users with the cost of limited user interaction. To ensure data integration and good usability we have developed the core data manipulation and storing functions using Web based technology and for data analysis we are using robust applications.

2. Application interfaces

Excellent usability does not only account for primely data interfaces. The ability to easily import data and the availability of well defined application interfaces are also crucial. Different institutions use diverse, mostly self tailored

applications with proprietary and varying data formats. MARS provides several data and application interfaces. To import data we provide user definable and manageable parsers. When a user is uploading data, MARS tries to find an appropriate parser based on the file data or format header. Once the data is uploaded and stored, the data can be analyzed using the provided applications. For scientists who would like to analyze their data with other software, MARS provides also a Web service data interface. After some slight adaptations, users can authenticated and down- or upload data. Providing a Web service interface allows through its wide spread and platform independence to be implemented in all well-established programming languages and in tools like Matlab or BioConductor.

Existing Web applications can be plugged-in using the EACI that enables the linkage between data provided by the plugged-in application and data stored in MARS. Moreover it is possible to extend MARS without having to amend the MARS source code.

3. Quality management

In order to assure high-quality data and to understand or optimize lower value data it is important to be able to trace back all conducted quality control steps. MARS traces several quality measurements performed during the microarray production as well as during the sample preparation, extraction, and hybridization process. These quality checks are implemented as an additional application called MARS-QM, which is tightly integrated into MARS.

4. Data sharing and export

MARS enables users to share their datasets with other users. Supplementary to the user oriented data management an institution oriented level has been introduced. This amelioration allows several institutes to store their data into one data repository without having to share common settings and resources such as scanners, but offering the possibility to share the data among them.

Besides the sharing of microarray experiment data we provide the possibility to export hybridizations and experiments using the common exchange format MAGE-ML. This feature facilitates the easy sharing and publishing of high quality, well annotated data within the life sciences community by uploading the generated files to public repositories like ArrayExpress [26].

5. User management

Since microarray- as well as the corresponding quality control data may contain highly sensitive data, we have integrated our AAS into MARS to provide authentication and fine grained authorization mechanisms. The combination of AAS and External Application Connector Interface provides through a single-sign-on mechanisms and dynamic linkage of data the possibility to assemble heterogeneous Web applications to one powerful suite.

Because information attached to molecules is changing quickly, we are currently implementing the possibility to update and enhance the information tagged to a molecule. Changing this information on the molecule level may affect already existing results. In order to avoid such precarious alterations, a user should be able to update the molecule information for each experiment separately instead of replacing the initial molecule information. Further ongoing projects concentrate on the integration of Affymetrix GeneChip arrays into MARS and the improvement of MAGE-ML export capabilities in order to obtain approval from the ArrayExpress annotation team. Both

features will be made available to the public in the next major release.

Conclusion

In summary, we have developed an integrated system consisting of a microarray database and a microarray quality control database, that has been tailored to serve the specific needs of microarray based research projects. Due to the unique fusion of using Web based and standalone applications connected to the latest J2EE application server technology, bioinformatics researchers receive the benefits of standards-based software engineering. The system can provide a model how to build up a similar platform for other emerging functional genomics technologies.

Availability and requirements

- Project name: MARS
- Project home page: <http://genome.tugraz.at/Software/MARS/MARS.html>
- Operating system: Solaris, Linux, Windows
- Programming language: Java, HTML
- Other requirements: Java JDK 1.4.x, Oracle 9i, MySQL 4.0.xx, Server with at least 1 GBytes of main memory
- License: IGB-TUG Software License
- Any restrictions to use by non-academics: no

Installation of MARS is not complicated and should be manageable within a few hours if necessary access rights especially to Oracle and MySQL are granted. Step-by-step instructions are provided at the projects Web site together with the files and scripts necessary for installation. The reference installation of MARS is running on a Sun Fire V880 server under Solaris 9 using Oracle 9i as Database Management System. Attached is a Storage Area Network (SAN) with 2 TBytes.

The production instance of MARS contains information from more than 1000 microtiter plates, 24 array batches, 232 hybridizations, and 312 rawbioassays with about 9,170,000 datapoints.

Authors' contributions

MM, RM, and AS designed and implemented the current version of MARS. They were responsible for the database design, the development of the business- as well as presentation logic. JH developed the quality management system and incorporated it into MARS. MM and JH were the lead developers of the AAS. HH, AP, GS, and MS have

been involved in the compilation of the user requirements document and contributed to the conception and design of the system. ZT was responsible for the overall conception and project coordination. All authors gave final approval of the version to be published.

Acknowledgements

The authors thank the staff of the Institute for Genomics and Bioinformatics for valuable comments and contributions. This work was supported by the Austrian Science Fund (Grant SFB Biomembranes F718) and the bm:bwk, GEN-AU BIN (Bioinformatics Integration Network) and GEN-AU GOLD (Genomics of Lipid-Associated Disorders). Michael Maurer, Robert Molitor and Juergen Hartler were supported by a grant from the Austrian Academy of Sciences.

References

- Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3**:RESEARCH0062.1-RESEARCH0062.12.
- Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci U S A* 1996, **93**:10614-10619.
- Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467-470.
- Haab BB, Dunham MJ, Brown PO: **Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions.** *Genome Biol* 2001, **2**:RESEARCH0004.1-RESEARCH0004.13.
- Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409**:533-538.
- Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23**:41-46.
- Yan H, Park SH, Finkelstein G, Reif JH, LaBean TH: **DNA-templated self-assembly of protein arrays and highly conductive nanowires.** *Science* 2003, **301**:1882-1884.
- Mousses S, Caplen NJ, Cornelison R, Weaver D, Basik M, Hautaniemi S, Elkahlon AG, Lotufo RA, Choudary A, Dougherty ER, Suh E, Kallioniemi O: **RNAi Microarray Analysis in Cultured Mamalian Cells.** *Genome Res* 2003, **13**:2341-2347.
- Hessner MJ, Wang X, Khan S, Meyer L, Schlicht M, Tackes J, Datta MW, Jacob HJ, Ghosh S: **Use of a three-color cDNA microarray platform to measure and control support-bound probe for improved data quality and reproducibility.** *Nucleic Acids Res* 2003, **31**:e60-e60.
- Tsangaris GT, Botsonis A, Politis I, Tzortzatou-Stathopoulou F: **Evaluation of cadmium-induced transcriptome alterations by three color cDNA labeling microarray analysis on a T-cell line.** *Toxicology* 2002, **178**:135-160.
- MGED – Microarray Gene Expression Data Society Home Page** [<http://www.mged.org>]
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
- Stoeckert CJ Jr, Causton HC, Ball CA: **Microarray databases: standards and ontologies.** *Nat Genet* 2002, **32(Suppl)**:469-473.
- Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Jordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A: **Design and implementation of microarray gene expression markup language (MAGE-ML).** *Genome Biol* 2002, **3**:RESEARCH0046.1-RESEARCH0046.9.
- Gene C Ontology: **Creating the gene ontology resource: design and implementation.** *Genome Res* 2001, **11**:1425-1433.
- Pasquier C, Girardot F, Jevardat dFK, Christen R: **THEA: ontology-driven analysis of microarray data.** *Bioinformatics* 2004, **20**:2636-2643.
- Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z: **PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.** *Nucleic Acids Res* 2005 in press.
- Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet.** *Nucleic Acids Res* 2002, **30**:42-46.
- BioCarta – Charting Pathways of Life** [<http://www.biocarta.com>]
- Kokocinski F, Wrobel G, Hahn M, Lichter P: **QuickLIMS: facilitating the data management for DNA-microarray fabrication.** *Bioinformatics* 2003, **19**:283-284.
- Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3**:SOFTWARE0003.1-SOFTWARE0003.6.
- Gardiner-Garden M, Littlejohn TG: **A comparison of microarray databases.** *Brief Bioinform* 2001, **2**:143-158.
- Killion PJ, Sherlock G, Iyer VR: **The Longhorn Array Database (LAD): An Open-Source, MIAME compliant implementation of the Stanford Microarray Database (SMD).** *BMC Bioinformatics* 2003, **4**:32-32.
- Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, Schroeder M, Brown PO, Botstein D, Sherlock G: **The Stanford Microarray Database: data access and quality assessment tools.** *Nucleic Acids Res* 2003, **31**:94-96.
- Theilhaber J, Ulyanov A, Malanchara A, Cole J, Xu D, Nahf R, Heuer M, Brockel C, Bushnell S: **GECKO: a complete large-scale gene expression analysis platform.** *BMC Bioinformatics* 2004, **5**:195-195.
- Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA: **ArrayExpress – a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31**:68-71.
- Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30**:207-210.
- JBoss.com: The Professional Open Source Company** [<http://www.jboss.org>]
- Quackenbush J: **Data standards for 'omic' science.** *Nat Biotechnol* 2004, **22**:613-614.
- Dudoit S, Fridlyand J: **Bagging to improve the accuracy of a clustering procedure.** *Bioinformatics* 2003, **19**:1090-1099.
- The MathWorks – Matlab and Simulink for Technical Computing** [<http://www.mathworks.com>]
- Samba – opening windows to a wider world** [<http://www.samba.org>]
- OpenLDAP** [<http://www.openldap.org>]
- Pieler R, Sanchez-Cabo F, Hackl H, Thallinger GG, Trajanoski Z: **ArrayNorm: comprehensive normalization and analysis of microarray data.** *Bioinformatics* 2004.
- Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data.** *Bioinformatics* 2002, **18**:207-208.
- Trost E, Hackl H, Maurer M, Trajanoski Z: **Java editor for biological pathways.** *Bioinformatics* 2003, **19**:786-787.

PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways

Bernhard Mlecnik, Marcel Scheideler, Hubert Hackl, Jürgen Hartler, Fatima Sanchez-Cabo and Zlatko Trajanoski*

Institute for Genomics and Bioinformatics and Christian-Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, Graz 8010, Austria

Received February 14, 2005; Revised and Accepted March 11, 2005

ABSTRACT

While generation of high-throughput expression data is becoming routine, the fast, easy, and systematic presentation and analysis of these data in a biological context is still an obstacle. To address this need, we have developed PathwayExplorer, which maps expression profiles of genes or proteins simultaneously onto major, currently available regulatory, metabolic and cellular pathways from KEGG, BioCarta and GenMAPP. PathwayExplorer is a platform-independent web server application with an optional standalone Java application using a SOAP (simple object access protocol) interface. Mapped pathways are ranked for the easy selection of the pathway of interest, displaying all available genes of this pathway with their expression profiles in a selectable and intuitive color code. Pathway maps produced can be downloaded as PNG, JPG or as high-resolution vector graphics SVG. The web service is freely available at <https://pathwayexplorer.genome.tugraz.at>; the standalone client can be downloaded at <http://genome.tugraz.at>.

INTRODUCTION

The huge amount of large-scale gene and protein expression data requires new methods for correlation with prior biological knowledge. Intense efforts have been undertaken and a variety of different computational methods have been applied to analyze these data sets. Initial data analyses are based on tools (1) using common clustering algorithms (2–6). Subsequent analyses require tools for visualizing genes and gene products in the context of biological pathways.

In the past years, scientists have established common repositories for biological data to make it freely available to others. Consequently, relating biological data to relevant pathway overviews is now accessible through publicly available resources. These sources include large graphical collections with open source access [e.g. Kyoto Encyclopedia of Genes and Genomes KEGG (7) at <http://www.genome.jp/kegg>, BioCarta at <http://biocarta.com/genes> and GenMAPP at <http://www.genmapp.org> (8)]. By overlaying expression data on biological pathways, established and novel relationships among genes can be explored. These pathways give key information about the functional and metabolic organization of cellular and biological systems within organisms.

Scientists working with large-scale expression data have already tried to verify and visualize their data in the context of biological pathways, but the analysis is hampered by a lack of systematic approaches. Growing trends of using freely available software packages in bioinformatics along with the increasing diversity of different operating systems are more and more claiming for platform-independent web solutions which are able to meet these requirements. Although some tools are available (9–13), limitations in performance, usability and functionality still remain. A user-friendly graphical user interface is critical for optimal usability by a broad range of users. Furthermore, there should be several methods available for downloading user-relevant information in a textual and graphical way.

For this purpose, we have developed a web-based service PathwayExplorer, which provides comprehensive and easily accessible representations of expression profiles onto major regulatory, metabolic and cellular pathways. The integrated pathway resources include KEGG, BioCarta and GenMAPP.

METHODS

We used state-of-the-art Java technologies to develop PathwayExplorer. It is an entirely Java-based application

*To whom correspondence should be addressed. Tel: +43 316 873 5332; Fax: +43 316 873 5340; Email: zlatko.trajanoski@tugraz.at

client using a three-tiered-architecture that ensures a clean separation between the presentation front-end, business and database back-end layer. The business layer, a Java application client conforming to the J2EE specification, performs the calculations and search functions and can be accessed by the presentation layer in two ways: (i) an application client using SOAP (simple object access protocol); and (ii) a web browser-based application client running on a web server using JSP technology. The database layer called PathwayDB is based on an Oracle DBMS (data base management system), which is also portable to freely available MySQL or PostgreSQL DBMS. Frequently changing information is kept in flat files, which are obtained and constantly updated from NCBI (<ftp://ftp.ncbi.nih.gov/refseq/LocusLink/>) (14) and KEGG (<ftp://ftp.genome.ad.jp/pub/kegg/ligand/>) (7).

PathwayDB minimizes the ambiguity among its gene identifiers. The fact that almost all identifiers are relationally and hierarchically linked allows it to specify the gene element nodes with only one kind of identifier, which constitutes the top of the hierarchical identifier tree. All the gene identifiers that lie below the root identifier can be linked to it later by using external data sources. We have successfully integrated KEGG, BioCarta and GenMAPP pathways into PathwayDB by using only the minimum information necessary. This comprises information from parsed SMBL (Systems Biology Markup Language) (15) files obtained from KEGG, which were converted into the PathwayExplorer application client (16) format. This was performed because of the lack of the SMBL format for encoding feasible graphical visualization, which is essential for the graphical evaluation of mapped pathways. The EC (17) (Enzyme Commission) defines the root identifier, which can hold several gene identifiers from all available organisms, i.e. the LocusLinks (they can contain again several gene identifiers, such as RefSeq or UniGene IDs) or the official gene identifiers for other organisms. To integrate BioCarta and GenMAPP into PathwayDB, the PathwayExplorer application client was once again used for automatically parsing the HTML pages holding the necessary pathway information.

Since both of these pathway resources use many different gene identifiers, LocusLink was again used as root identifier. The LocusLinks are linked with the user-defined gene identifier groups (UniGene, GeneOntology, GenBank and/or RefSeq), which are used then to align the mapped gene IDs.

PROGRAM DESCRIPTION

Accessibility

PathwayExplorer is a web-based service constantly available at <https://pathwayexplorer.genome.tugraz.at>, with a public, login-free data repository for uploading data sets.

The PathwayExplorer standalone client application can perform the same mapping operations on an independent, local-platform computer system. In this case, instead of uploading the expression data to the web server, the pathway information from PathwayDB is downloaded to the user's local computer system. The standalone client connects to PathwayDB through a SOAP interface. The standalone client is available at the PathwayExplorer homepage or at <http://genome.tugraz.at/Software/PathwayExplorer/Setup.html>.

Table 1. The number of unique gene identifiers (e.g. *Homo sapiens*) available for mapping expression profiles in PathwayExplorer

Pathway resource	No. of pathways	Unique RefSeq accession no.	Unique GenBank accession no.	Unique UniGene accession no.	Unique GO accession no.
KEGG	120	4099	26589	2827	2561
BioCarta	311	2209	15889	1438	1671
GenMAPP	82	6374	38171	4527	2857
Sum	513	8947	55111	6276	3623

The table shows the sum of non-redundant accession numbers available.

Input

As input, PathwayExplorer receives a common tab-delimited text file containing expression profiles with the gene identifier as first column, the gene name as an optional second column and any experiment or time point data as further columns. Possible gene identifiers for organisms using LocusLinks are GenBank accession numbers, RefSeq IDs, UniGene IDs and Gene Ontology IDs (e.g. see Table 1). For all other organisms, systematic gene identifiers are possible. The RefSeq IDs are used as the initial default gene identifier group, and this can be changed later. The uploaded gene-expression data sets can be stored in either a public or a login-requiring repository where they can be modified or deleted again.

Calculations and visualization

An example for mapping a public data set from a yeast sporulation study (18) is given in Figure 1. In order to map data sets onto pathways, the user is requested to select the organism and the data set to be mapped. The loaded data set remains in the background as long as it has not been closed, and subsequently every pathway which becomes opened is then automatically mapped with this data set. To restrict the uploaded data set to certain criteria before mapping (e.g. to use only expression profiles of differentially expressed gene or proteins), filter options can be applied: (i) to filter out expression profiles with too many missing data points; (ii) to filter out weakly expressed profiles (based on a certain standard deviation threshold); and (iii) to filter out genes whose expression values do not meet a certain threshold.

After filtering the data set, PathwayExplorer provides two mapping options: (i) mapping the data set to a single pathway by choosing one in the hierarchical tree or (ii) Mapping the data set onto all available pathways at once. Option (ii) generates a list (Figure 3), which ranks all mapped pathways by their number of mapped genes and allows for sorting the list based on different criteria, such as (a) pathway name; (b) unique gene identifiers available in each pathway; (c) the number of gene identifiers which has passed the filter criteria and was mapped to the pathway; (d) the number of genes which would have been mapped to the pathway if they had passed the filter criteria; and (e) the right-tailed *P*-value of a Fisher's exact test (f) the false discovery rate (FDR) (19) corrected *Q*-value.

With a right-tailed Fisher's exact test, we test whether the proportion of mapped genes within the set of differentially expressed genes is significantly larger than the proportion of

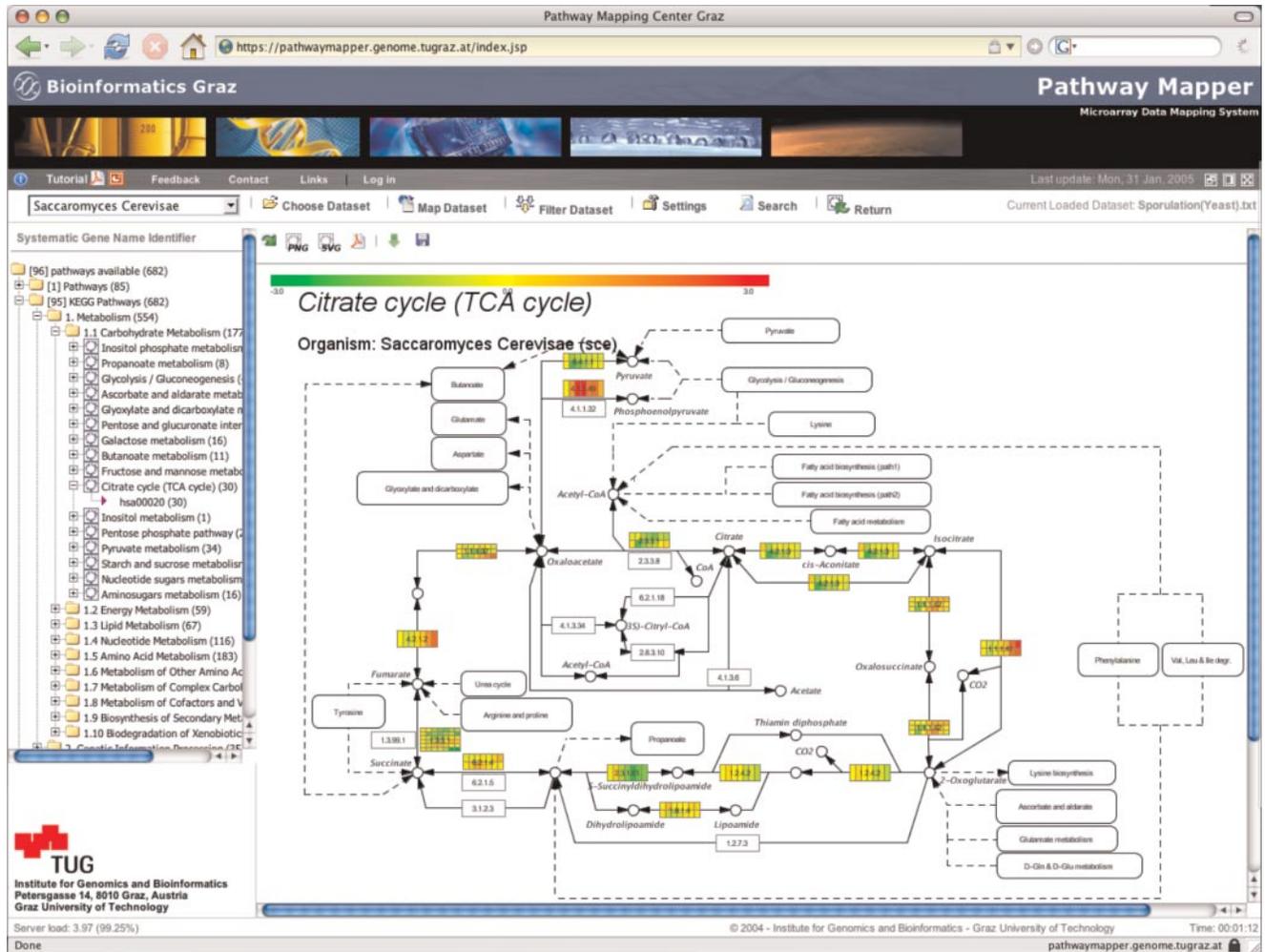


Figure 1. PathwayExplorer example: a screenshot of a pathway mapped with expression data. (i) The toolbar frame (the row including the organism field) offers various setup and visualization options. (ii) Hierarchical tree frame (on the left) enables browsing through all available pathway sections. (iii) The main frame (in the center) displays the citrate cycle pathway extracted from KEGG, which mapped with a yeast sporulation data set with seven different time points (0, 0.5, 2, 5, 7, 8 and 11.5 h). The color-coded boxes represent mapped genes. If more than just one gene ID (e.g. RefSeq) matches to a box, each box is split up into several horizontal elements. According to the number of experiments/time points (in this case seven time points), the boxes are split again into vertical columns which display the expression level of each time point. To visualize a mapped expression profile in a different way, one has to choose the corresponding horizontal row of a box (see Figure 2).

Table 2. 2 × 2 contingency table for the Fisher’s exact test

	Genes that are differentially expressed (passed the filter)	Gene that are not differentially expressed (filtered out)	
Mapped genes	A	B	A+B
Unmapped genes	C	D	C+D
	A+C	B+D	Total number of genes

The null hypothesis of the right-tailed Fisher’s exact test states that the proportion of A/C is smaller or equal to the proportion of B/D. If the right-tailed P-value is <5%, we reject the null hypothesis, which means that the proportion of differentially expressed genes is significantly greater than those that are not differentially expressed.

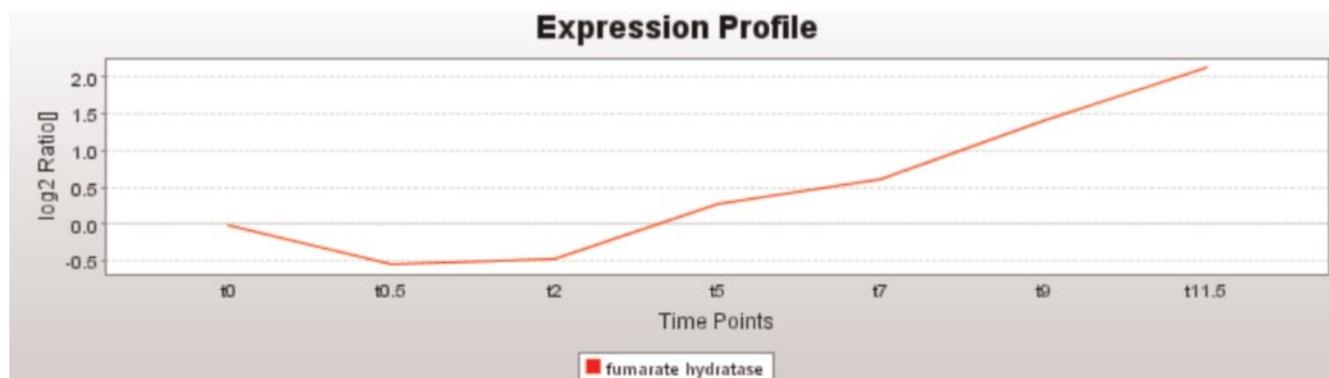
mapped genes that are not differentially expressed (see Table 2). We use a Fisher’s exact test because the number of counts might be smaller than five for any of the fields in the contingency table. Multiple hypotheses correction (19) is needed to

control the number of false positives, since many hypotheses are tested simultaneously.

The graphical visualization of the displayed pathway image can be changed to the SVG view using the freely available SVG Viewer plug-in from Adobe (<http://www.adobe.com>). This enables on-line zooming of the pathway graphic.

Output

PathwayExplorer provides graphical and textual output. It generates a gene cluster for each mapped pathway, which can be downloaded in the same tab-delimited text format as the uploaded data set. Each mapped expression profile can be displayed (Figure 2) by selecting the corresponding box (Figure 1) on the pathway image. The generated ranking list for all mapped pathways (Figure 3) can also be downloaded as tab-delimited text file and can be used for statistical analyses.



GENE	
Symbol Name	4.2.1.2
Gene Name	fumarate hydratase
Accession number (RefSeq)	YPL262W

Figure 2. By selecting one row (if there are more than one) of a mapped gene box (Figure 1), the corresponding gene and expression profile information will be displayed. To obtain additional information links to GenBank, Entrez and OMIM are provided. Only one mapped expression profile from the uploaded data set can be displayed at once.

STATISTICS	Passed UniqlDs	Filtered out UniqlDs	Σ
Mapped to Pathways	357	244	601
Not Mapped to Pathways	2617	2882	5499
Σ	2974	3126	6100

No.	Id	Section	Subsection	Pathway	Pathway UniqlD	Passed UniqlD	Passed UniqlD %	Filtered UniqlD	p-value	q-value
Σ				104	684	357	52.19	244	20	7
1	hsa00710	KEGG Pathways	1.2 Energy Metabolism	Carbon fixation	18	14	77.78	2	0.0020	0.037
2	hsa00630	KEGG Pathways	1.1 Carbohydrate Metabolism	Glyoxylate and dicarboxylate metabolism	15	12	80.0	1	0.0010	0.037
3	hsa00051	KEGG Pathways	1.1 Carbohydrate Metabolism	Fructose and mannose metabolism	15	12	80.0	1	0.0010	0.037
4	hsa00251	KEGG Pathways	1.5 Amino Acid Metabolism	Glutamate metabolism	27	21	77.78	4	0.0	0.037
5	hsa00230	KEGG Pathways	1.4 Nucleotide Metabolism	Purine metabolism	95	56	58.95	28	0.0010	0.037
6	hsa00252	KEGG Pathways	1.5 Amino Acid Metabolism	Alanine and aspartate metabolism	29	21	72.41	6	0.0020	0.039
7	hsa00010	KEGG Pathways	1.1 Carbohydrate Metabolism	Glycolysis / Gluconeogenesis	47	28	59.57	10	0.0020	0.047
8	hsa00530	KEGG Pathways	1.1 Carbohydrate Metabolism	Aminosugars metabolism	16	13	81.25	2	0.0030	0.056

Figure 3. (a) Shows the overall statistic of the filtered unique identifiers of the expression data set mapped on all pathways. (b) The ranking list of all mapped pathways is also displayed and can be sorted by different criteria, allowing easy navigation through all pathways. By selecting the pathway of interest, the data set will be automatically mapped on the pathway and the expression values will be displayed in a selectable and intuitive color code (see Figure 1). The last two columns display the *P*-value and the FDR (19) corrected *Q*-value.

A special feature is the PDF generator, which can be applied for every single pathway, irrespective of whether the genes were mapped to the pathway or not. This generator creates a PDF document of the currently loaded pathway, which can be downloaded or directly displayed in the user's web browser. If genes were mapped to the pathway, the PDF generator additionally adds the expression profiles to the document. In the case of human expression data sets, additional information about each gene is directly extracted from the OMIM (Online Mendelian Inheritance in Man) database. This feature offers the user a special opportunity to get a quick and comprehensive overview of the current pathway plus detailed information about each mapped gene. The graphical output of PathwayExplorer can be directly downloaded in PNG or SVG graphic format.

CONCLUSION

We have developed PathwayExplorer, a web server providing comprehensive and facile mapping of gene or protein expression profiles. The profiles are simultaneously mapped onto the major regulatory, metabolic and cellular pathways available from the KEGG, BioCarta and GenMAPP pathway resources. The server accepts expression data files in a tab-delimited text format and generates high-resolution vector graphic images of mapped pathways. It enables further very compact representations of expression profiles within all available pathways. PathwayExplorer not only unifies the access to different pathway resources, but also combines gene identifiers arbitrarily selectable by the user.

PathwayExplorer is at present limited by common gene identifiers, such as RefSeq, GenBank, Gene Ontology or UniGene, but due to PathwayExplorer's flexible design, future requirements can be easily integrated. PathwayExplorer is thus endowed with a wide range of functionality giving the user multiple options to extract biological information in a comprehensive systematic and intuitive way. The online accessibility and the intuitive interface of PathwayExplorer should make it a valuable tool for a broad range of users.

ACKNOWLEDGEMENTS

The authors thank all members of the Institute for Genomics and Bioinformatics and external scientists who contributed to this work and James McNally for critically reviewing the manuscript. This work was supported by the Austrian Science Fund (Grant SFB Biomembranes F718) and the bm:bwk, GEN-AU:BIN, Bioinformatics Integration Network. B.M. was supported by a grant from the Austrian Academy of Sciences (OEAW). Funding to pay the Open Access publication charges for this article was provided by GEN-AU:BIN.

Conflict of interest statement. None declared.

REFERENCES

1. Sturn,A., Mlecnik,B., Pieler,R., Rainer,J., Truskaller,T. and Trajanoski,Z. (2003) Client-server environment for high-performance gene expression data analysis. *Bioinformatics*, **19**, 772–773.
2. Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
3. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
4. Wen,X., Fuhrman,S., Michaels,G.S., Carr,D.B., Smith,S., Barker,J.L. and Somogyi,R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 334–339.
5. Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M., Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
6. Fellenberg,K., Hauser,N.C., Brors,B., Neutzner,A., Hoheisel,J.D. and Vingron,M. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.
7. Kanehisa,M., Goto,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
8. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.*, **31**, 19–20.
9. Pandu,R., Guru,R.K. and Mount,D.W. (2004) Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, **20**, 2156–2158.
10. Chung,H.J., Kim,M., Park,C.H., Kim,J. and Kim,J.H. (2004) ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res.*, **32**, W460–W464.
11. Sirava,M., Schäfer,T., Eiglsberger,M., Kaufmann,M., Kohlbacher,O., Bomber-Bauer,E. and Lenhof,HP. (2002) BioMiner—modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*, **18**, 219–230.
12. Goesmann,A., Haubrock,M., Meyer,F., Kalinowski,J. and Giegerich,R. (2002) PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, **18**, 124–129.
13. Pan,D., Sun,N., Cheung,K.H., Guan,Z., Ma,L., Holford,M., Deng,X. and Zhao,H. (2003) PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. *BMC Bioinformatics*, **32**, W460–W464.
14. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
15. Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
16. Trost,E., Hackl,H., Maurer,M. and Trajanoski,Z. (2003) Java editor for biological pathways. *Bioinformatics*, **19**, 786–787.
17. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.
18. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
19. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Database

Open Access

GOLD.db: genomics of lipid-associated disorders database

Hubert Hackl, Michael Maurer, Bernhard Mlecnik, Jürgen Hartler, Gernot Stocker, Diego Miranda-Saavedra and Zlatko Trajanoski*

Address: Institute for Genomics and Bioinformatics and Christian-Doppler-Laboratory for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria

Email: Hubert Hackl - hubert.hackl@tugraz.at; Michael Maurer - michael.maurer@tugraz.at; Bernhard Mlecnik - bernhard.mlecnik@tugraz.at; Jürgen Hartler - juergen.hartler@tugraz.at; Gernot Stocker - gernot.stocker@tugraz.at; Diego Miranda-Saavedra - diego@compbio.dundee.ac.uk; Zlatko Trajanoski* - zlatko.trajanoski@tugraz.at

* Corresponding author

Published: 10 December 2004

Received: 06 September 2004

BMC Genomics 2004, 5:93 doi:10.1186/1471-2164-5-93

Accepted: 10 December 2004

This article is available from: <http://www.biomedcentral.com/1471-2164/5/93>

© 2004 Hackl et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The GOLD.db (Genomics of Lipid-Associated Disorders Database) was developed to address the need for integrating disparate information on the function and properties of genes and their products that are particularly relevant to the biology, diagnosis management, treatment, and prevention of lipid-associated disorders.

Description: The GOLD.db <http://gold.tugraz.at> provides a reference for pathways and information about the relevant genes and proteins in an efficiently organized way. The main focus was to provide biological pathways with image maps and visual pathway information for lipid metabolism and obesity-related research. This database provides also the possibility to map gene expression data individually to each pathway. Gene expression at different experimental conditions can be viewed sequentially in context of the pathway. Related large scale gene expression data sets were provided and can be searched for specific genes to integrate information regarding their expression levels in different studies and conditions. Analytic and data mining tools, reagents, protocols, references, and links to relevant genomic resources were included in the database. Finally, the usability of the database was demonstrated using an example about the regulation of Pten mRNA during adipocyte differentiation in the context of relevant pathways.

Conclusions: The GOLD.db will be a valuable tool that allow researchers to efficiently analyze patterns of gene expression and to display them in a variety of useful and informative ways, allowing outside researchers to perform queries pertaining to gene expression results in the context of biological processes and pathways.

Background

The excessive consumption of high calorie, high fat diets and the adoption of a sedentary life style have made obesity and atherosclerosis major health problems in Western societies. In the USA, over 50% of the population are overweight (BMI > 25) and close to 25% are considered obese

(BMI > 30) [1,2]. As a consequence, a large fraction of the population is at risk to develop a broad range of common, life-threatening diseases including non-insulin dependent diabetes, various hyperlipidemias, high blood pressure and atherosclerosis. Vascular disease including coronary heart disease and stroke is currently the major cause of

death in the United States and in other industrialized nations.

At the root of obesity and atherosclerosis is an excessive deposition of neutral lipids. Adipose tissue accumulates predominantly triglycerides, whereas macrophages along the blood vessel wall mainly accumulate cholesterol and cholesteryl esters. Accordingly, a detailed understanding of the molecular mechanisms that govern the balance between lipid deposition and mobilization is fundamentally important for the prevention and improved treatment of disease. In addition to the apparent environmental components involved in the pathogenesis of disorders related to lipid and energy metabolism, a large number of studies have provided undisputed evidence that susceptibility genes contribute around 50% of the phenotype. These genes encode products involved in the cellular uptake, synthesis, deposition and/or mobilization of lipids. However, characterization of many if not most of these genes and their products remains rudimentary. Deficiencies in the current level of understanding extend to key enzymes such as important triglyceride hydrolases in adipose tissue [3] or cholesteryl ester hydrolases in macrophages, hormones, signal transduction pathways, and the regulation of the transcription of relevant genes.

While medical molecular biology traditionally associates single genes and gene products with diseases, a growing body of evidence suggests that several common disease phenotypes arise from the delicate interaction of many genes as well as gene-environment interactions. To elucidate the development of obesity and atherosclerosis, it will be necessary to analyze patterns of gene expression and relate them to various metabolic states. To discover novel genes, processes and pathways that regulate lipid deposition and mobilization, a departure from hypothesis-driven research and turn to a discovery-driven approach is necessary. The application of high-throughput technologies and genome-based analysis will provide the tools for the analysis of gene-gene and gene-environment interactions in a systematic and comprehensive manner.

To facilitate genomic research we have initiated the development of a system for storing, integrating, and analyzing relevant data needed to decipher the molecular anatomy of lipid associated disorders. In order to provide a reference for pathways and information of the relevant genes and proteins in an efficiently organized way, we have created the Genomics Of Lipid-Associated Disorders database (GOLD.db). The GOLD.db integrates disparate information on the function and properties of genes and their protein products that are particularly relevant to the

biology, diagnosis management, treatment, and prevention of lipid-associated disorders.

Construction and content

The main goal of the GOLD.db was to provide biological pathways with image maps and visual pathway information. For each element in the pathway, specific information exists including structured information about a gene, protein, function, literature, and links. The GOLD.db provides also the possibility to map gene expression data individually to each pathway. Additionally, analytic and data mining tools, reagents, protocols, references, and links to relevant genomic resources were included in the database.

The GOLD.db was implemented in Java technology [4]. Hence, the pathway editor, as well the web application are platform independent. The web application of GOLD.db is built in Java Servlets and JavaServer Pages technology based on the Model-View-Controller Architecture [5]. For the implementation, the freely available struts framework [6] was used. This code can be easily deployed in any Servlet Container. We used the Servlet Container Tomcat (also freely available at [7]) which is accessible from all web browsers. Oracle 9i was used as database management system. The interface between the Java and the Database management system was established using Java database connectivity (JDBC) 2.0. Therefore, migration to other freely available DBMSs like MySQL can be easily done. For additional storage and communication between the pathway-editor components, the markup language XML containing structured, human readable information, was used. The provided pathways can be downloaded as Scalable Vector Graphics (SVG) [8], a standard for describing two-dimensional graphics in XML, and can be visualized in this format on the client side with the web browser using a plug-in for SVG.

For tracking the repository of the reagents like clone resources and libraries which can be used for microarray studies, we have developed a relational database. Information about the vector, the sequence and length of the clone insert, primers for the PCR amplification, tissue, organism, accession number, library, container, storage information, date and person and access to other clone bases (e.g. IMAGE Consortium) can be stored. Users of the GOLD.db can list these clones and get all the information about each available clone. With restricted access, clone information or even clone lists can be uploaded and selection lists can be created and deleted. The input mask is designed in such way that the user can choose one of the elements of the created selection lists.

In order to deal with the huge amount of data associated with large scale studies and to perform sequence based

and microarray analysis, several bioinformatic tools were integrated or can be downloaded. Sequence similarity search against databases can be performed with BLAST (Basic Local Alignment Search Tool) [9], FASTA [10] or HMM (Hidden Markov Models) [11] on a 50 CPU Myrinet Cluster. The sequence retrieval system SRS (LION Bioscience AG, Heidelberg, Germany) was included to enable rapid, easy and user friendly access to the large volumes of diverse and heterogeneous data [12]. The latest version of the PathwayEditor for the construction of biological pathway diagrams can be downloaded. For microarray analysis the platform independent JAVA tools ArrayNorm [13] for normalization of microarray data and Genesis [14] for clustering and analysis of large scale gene expression datasets were made available.

Utility and discussion

Pathways

In order to construct the biological pathways of interest, we have developed a pathway editor [15] and an extended version to map gene expression data (pathway mapper). This drawing tool provides the possibility to draw elements – typically representing a gene as part of the pathway – and the connection between those elements. The benefit of this tool is that information can be appended to each element via an input mask. This information can be accessed by clicking on the corresponding element in the image map within the pathway mapper or when saved and uploaded via the web interface to the GOLD.db. To design this pathway service as flexible as possible, features are provided for the remove, up- and download of relevant pathways (image maps) including the underlying additional information of the elements. However, this service is on a restricted basis to prohibit unauthorized access. Since some pathways tend to become very detailed an option to search for genes or gene accession number, respectively, within the pathways was built in. The pathway editor is executable as a standalone application and is available from [16]. Currently annotated pathways are the insulin signaling pathway, the IGF-1 pathway and the adipogenesis regulatory network. Other pathways of lipid metabolism will follow in the near future. Available KEGG pathways can also be adapted with the pathway editor based on the provided XML files [17] and uploaded in the same way. All relevant KEGG pathways for different organisms are provided. Moreover, pathways from BioCarta were made available within the GOLD.db and HTML files [18] were parsed to provide additional meta-information of the pathway elements.

For each element in the pathways a specific information field exists. The field includes structured information about a gene, protein, function, literature, and links to well-curated and annotated databases. Besides the gene name and the symbol name – for human the HUGO sym-

bols and gene names and for mouse the MGI nomenclature were used – RefSeq numbers for the transcript and the protein as well as a link to SwissProt/UniProt and LocusLink is available. For the elements of the KEGG pathways a link to the provided enzyme or product information was given. The description, localization and classification of the factors are entered by the annotator in plain text and are accessed in the same format. The references used to generate the content of the database entries can be appended, including a link to the PubMed entry. There is also the possibility to create a list of reference entries for the pathway. If a clone for a specific gene is available in the clone resources, the clone name will be displayed automatically and a link with optional information about this clone is provided.

Mapping of gene expression data sets to pathways

Through the integration of several types of biological information deeper insights into the molecular mechanisms and biological processes can be gained than just by the analysis of one type of experimental results. In the GOLD.db it is possible to map gene expression data (for instance results of microarray studies) to the corresponding elements of the available pathways similar to previous efforts [19]. Either an individual or a provided gene expression data set can be used to visualize the gene expression at different experimental conditions sequentially or all at once in the context of a pathway. If an element (gene) of the pathway is included in the data set, the related symbol in the image map is color coded according to the relative gene expression or the log ratio in two color microarray experiments, respectively.

As key for the mapped relation the RefSeq number [20] is used. Hence, only those elements in the data set file are mapped, where the RefSeq number in the data set is specified. For the KEGG pathways each element classified by the enzyme classification number (EC) is virtually subdivided into different corresponding RefSeq entries, since one EC is represented by one or more RefSeq entries.

Curated gene expression data sets

Analysis of gene expression patterns in animal and cell models for lipid-associated disorders will help to understand the fundamental gene relations and regulatory mechanisms responsible for the development of obesity related diseases. The huge amount of data associated with the analysis of large scale gene expression analysis raises the demand of tools for storing, processing and retrieving complex information. Although a number of studies have been published and despite the requirements of some journals to deposit microarray data in public databases like GEO <http://www.ncbi.nlm.nih.gov/geo/> or ArrayExpress <http://www.ebi.ac.uk/arrayexpress/>, it is still very difficult for researchers to obtain the original data. Web

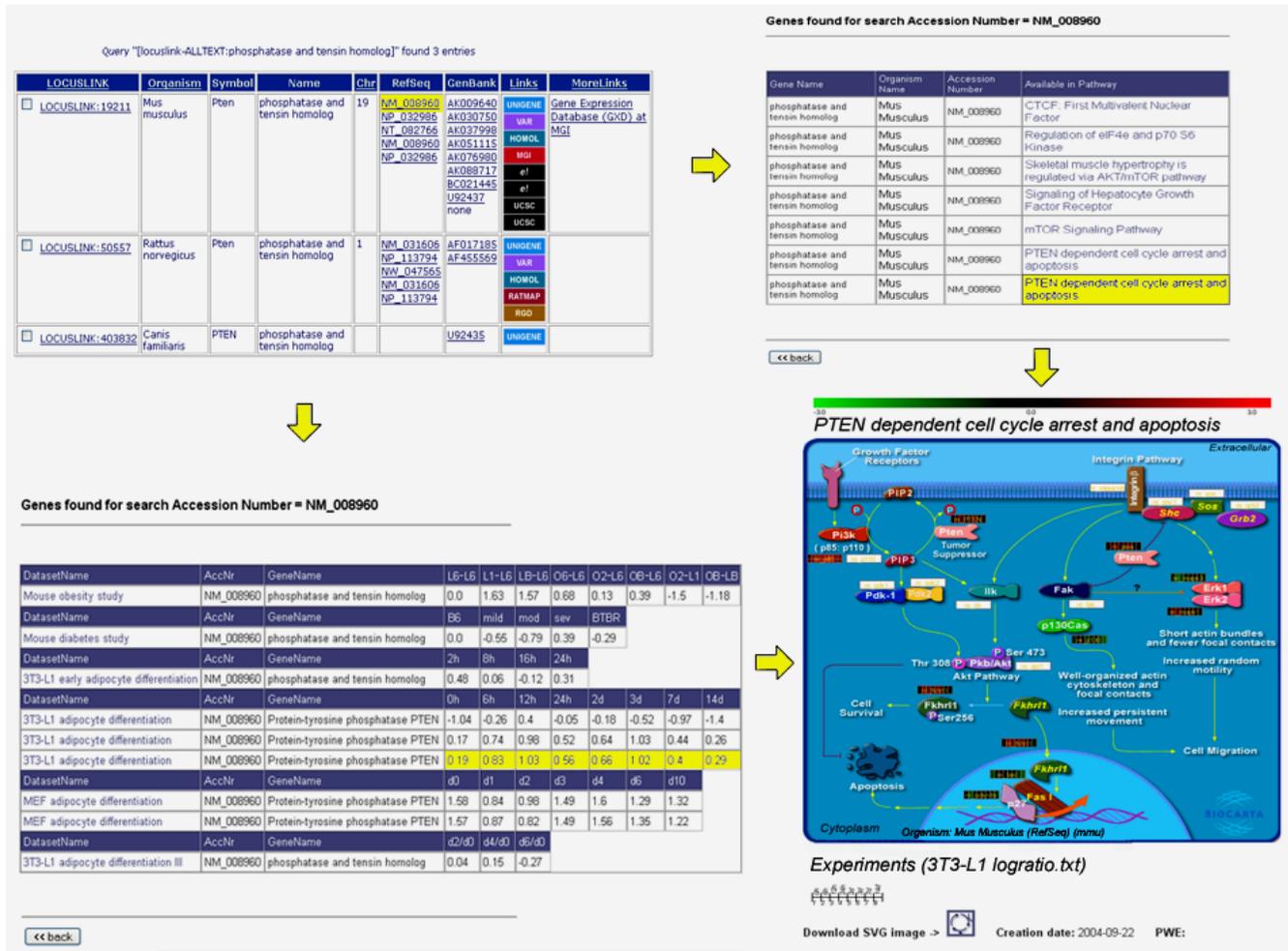


Figure 1 Various result tables from using GOLD.db to address the question how is PTEN regulated during adipocyte differentiation (*top left*: result of search in SRS for phosphatase and tensin homolog; *top right*: pathways, in which PTEN is involved; *bottom left*: relative gene expression levels of PTEN in different datasets; *bottom right*: PTEN dependent cell cycle pathway with mapped gene expression levels)

sites with Supplementary information are not maintained and/or not further developed. Hence, a database with a large collection of curated datasets will be enormously valuable for the community. Approaches to upload and retrieve gene expression data were pursued within the GOLD.db. Large scale gene expression data sets can be uploaded in form of tab delimited text files (Stanford file format) [21] as used for cluster analysis programs together with additional information about the experimental conditions and the citation for already published data sets. Within those data sets the search for specific genes is possible to provide integrated visualization of gene expression levels in different studies and experimental conditions.

Example for using GOLD.db: regulation of Pten during adipocyte differentiation

Recently, it was shown that insulin sensitivity, energy expenditure, and thermogenesis were enhanced in adipose-specific Pten-deficient (AdipoPten-KO) mice. Body and adipose tissues weight in these mice were significantly lower than those of control mice in spite of a larger food intake [22]. We addressed the question how is the expression of the Pten gene regulated during adipocyte differentiation in different models and experimental setups and in which pathways is PTEN involved. The workflow for the analysis is described in Figure 1. Pten (phosphatase and tensin homolog deleted on chromosome 10) is known as tumor suppressor gene and is a protein and lipid phos-

phatase with the major substrate phosphatidylinositol 3,4,5-triphosphate (PIP3), as indicated in the annotated insulin signaling pathway within the GOLD.db. In fact, Pten regulates negatively the insulin signaling pathway in 3T3-L1 adipocytes [23].

During adipocyte differentiation cyclin dependent kinase inhibitors, like p21 leads to a hypophosphorylation of the Retinoblastoma protein (Rb) which allows binding to the E2F transcription factor, causing cells to permanently exit the cell cycle – a required step in adipocyte differentiation called mitotic clonal expansion – before entering the terminal differentiation state. pRb interacts physically with adipogenic CCAAT/enhancer-binding proteins and positively regulates transactivation by C/EBP β and therefore plays a pivotal role in adipocyte differentiation [24,25]. Hence, since a) PTEN is expressed during adipogenesis (Figure 1), b) is involved in the regulation of Rb [22], a major player in adipogenesis, and c) is an important component in cell cycle arrest and apoptosis (Figure 1), it can be postulated that PTEN plays an important role in fat cell development.

Thus, using recently identified key player for food intake and weight control and using the GOLD.db, it is possible to address relevant questions and generate testable hypotheses on the molecular mechanisms of fat cell development.

Conclusions

The vast quantity of gene expression data generated in genomic studies presents a number of challenges for their effective analysis and interpretation. In order to fully understand the changes in expression that will be observed, we must correlate these data with phenotype, genotype, metabolism and other information including the tissue distribution and time course expression data gleaned from previous studies. The goal of our work was the development of a specialized database and tools that allow researchers to efficiently analyze patterns of gene expression and to display them in a variety of useful and informative ways, allowing outside researchers to perform queries pertaining to gene expression results in the context of biological processes and pathways. The uniqueness of the GOLDdb database we have developed is threefold: 1) the inclusion of annotated pathways, 2) the availability of curated datasets and 3) the possibility to map experimental data on biological pathways. The upcoming challenges will be to include data from functional analysis and proteomics data, which will give us new opportunities in understanding mechanisms of different applications and lipid-associated disorders in particular.

Availability and requirements

The GOLD.db database should be cited with the present publication as a reference. Access to GOLD.db is possible through the world wide web at <http://gold.tugraz.at>. The pathway editor and the clone tracker are available free of charge to academic, government, and other nonprofit institutions.

Author's contributions

HH was responsible for the content, the annotation process, webdesign, and processing of data sets. MM was responsible for the implementation of the database and web application as well as the relational database for the clone tracker. BM and JH had implemented the mapping of expression data to pathways. GS is involved in providing of sequence analysis tools and server software. DMS has annotated the insulin signaling pathway. ZT was responsible for the design of the study and for overall project coordination.

Acknowledgements

This work was supported by the Austrian Science Fund, Project SFB Biomembranes F718, the GEN-AU projects Bioinformatics Integration Network (BIN) and Genomics of Lipid-Associated Disorders (GOLD). Diego Miranda-Saavedra was supported by an EU Marie Curie Training Site program "Genomics of Lipid Metabolism". Michael Maurer was supported by a grant from the Austrian Academy of Sciences.

References

1. Flegal KM, Carroll MD, Kuczmarski RJ, Johnson CL: **Overweight and obesity in the United States: prevalence and trends, 1960-1994.** *Int J Obes Relat Metab Disord* 1998, **22**:39-47.
2. Must A, Spadano J, Coakley EH, Field AE, Colditz G, Dietz WH: **The disease burden associated with overweight and obesity.** *JAMA* 1999, **282**:1523-1529.
3. Zechner R, Strauss J, Frank S, Wagner E, Hofmann W, Kratky D, Hiden M, Levak-Frank S: **The role of lipoprotein lipase in adipose tissue development and metabolism.** *Int J Obes Relat Metab Disord* 2000, **24**:S53-S56.
4. **Java Technology** 2004 [<http://java.sun.com>].
5. Gamma E, Helm R, Johnson R, Vlissides J: *Design Patterns - Elements of Reusable Object-Oriented Software* Addison-Wesley; 1995.
6. **Struts framework** 2004 [<http://jakarta.apache.org/struts/>].
7. **Tomcat** 2004 [<http://jakarta.apache.org/tomcat/>].
8. **SVG** 2004 [<http://www.w3.org/TR/SVG>].
9. Altschul SF, Gish W, Miller W, Myers EV, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
10. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**:2444-2448.
11. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
12. Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks.** *Methods Enzymol* 1996, **266**:114-128.
13. Pieler R, Sanchez-Cabo F, Hackl H, Thallinger GG, Trajanoski Z: **ArrayNorm: comprehensive normalization and analysis of microarray data.** *Bioinformatics* 2004, **20**:1971-3.
14. Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data.** *Bioinformatics* 2002, **18**:207-208.
15. Trost E, Hackl H, Maurer M, Trajanoski Z: **Java editor for biological pathways.** *Bioinformatics* 2003, **19**:786-787.
16. **Institute for Genomics and Bioinformatics, Graz University of Technology** 2004 [<http://genome.tugraz.at>].
17. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet.** *Nucleic Acids Res* 2002, **30**:42-46.

18. **Biocarta Pathways** 2004 [<http://biocarta.com/genes/allPathways.asp>].
19. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.** *Nat Genet* 2002, **31**:19-20.
20. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29**:137-140.
21. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863-14868.
22. Komazawa N, Matsuda M, Kondoh G, Mizunoya W, Iwaki M, Takagi T, Sumikawa Y, Inoue K, Suzuki A, Mak TW, Nakano T, Fushiki T, Takeda J, Shimomura I: **Enhanced insulin sensitivity, energy expenditure and thermogenesis in adipose-specific Pten suppression in mice 2.** *Nat Med* 2004, **10**:1208-1215.
23. Nakashima N, Sharma PM, Imamura T, Bookstein R, Olefsky JM: **The tumor suppressor PTEN negatively regulates insulin signaling in 3T3-L1 adipocytes.** *J Biol Chem* 2000, **275**:12889-12895.
24. Hansen JB, Petersen RK, Larsen BM, Bartkova J, Alsner J, Kristiansen K: **Activation of peroxisome proliferator-activated receptor gamma bypasses the function of the retinoblastoma protein in adipocyte differentiation.** *J Biol Chem* 1999, **274**:2386-2393.
25. Chen PL, Riley DJ, Chen Y, Lee WH: **Retinoblastoma protein positively regulates terminal adipocyte differentiation through direct interaction with C/EBPs.** *Genes Dev* 1996, **10**:2794-2804.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

