Integration of Transcriptome and Proteome Data from Human-Pathogenic Fungi by Using a Data Warehouse

Daniela Albrecht*[1], Olaf Kniemeyer[1], Axel A. Brakhage[1], Matthias Berth[2], Reinhard Guthke[1]

[1]Leibniz-Institute for Natural Product Research and Infection Biology -
Hans-Knoell-Institute (HKI), Beutenbergstrasse 11a, 07745 Jena

[2]Decodon GmbH, BioTechnikum, Walther-Rathenau-Strasse 49a, 17489 Greifswald

Summary

A data warehouse for the integrated storage and visualisation of genome and experimental transcriptome and proteome data of human-pathogenic fungi was established. It provides tools for uploading images and corresponding data from microarray experiments, two-dimensional (2D) gel experiments and mass spectrometry (MS) analyses. All data are cross-linked. A user can find out, on which gels in the database an interesting protein was detected. Additionally, he can see on which microarrays the corresponding mRNA had been spotted and whether these spots show interesting intensity values. So the data warehouse enables an integrated analysis of both transcriptome and proteome data.
Some of the uploaded data were transcriptome and proteome time series data of temperature shift experiments obtained from *Aspergillus fumigatus*. Several proteins were differentially regulated at different times after the temperature shift. For a couple of them also the respective transcripts were found to be differentially expressed. For even more of those proteins the transcripts did not show differential regulation and vice versa. So both kinds of data clearly complement each other and should be analysed together.

1 Introduction

Today about 500 000 different variants of fungi have been described. Only about 100 of them are known to be human pathogens. During the last decades these fungi became menacing as opportunistic infectious microorganisms. Advances of modern medicine, including transplantations and other surgeries intensify this development. Additionally, the ageing of the population and the growing amount of patients undergoing immunosuppressive therapies result in a higher number of vulnerable persons [1]. In the majority of fungal infections patients are immunocompromised. This plays a major role in the type and course of the mycoses. From 1980 till 1990 the fraction of fungi causing nosocomial infections (i.e., infections acquired in hospitals) rose from 6 % to 11% within the USA [2]. The most common causes of such infections are *Candida* subspecies (in particular *Candida albicans*) and *Aspergillus* subspecies (mainly *Aspergillus fumigatus*). Both are facultative pathogenic fungi within the Deuteromycota, which are Ascomycota where no sexual life cycle is known. The infections caused by *Aspergillus*

*corresponding author, Daniela.Albrecht@hki-jena.de

species nearly quadrupled in the USA in the 1980s. In Germany there are about 40 000 serious invasive *Candida* infections per year and up to 10% of the patients die.
C. albicans is the most common fungal organism isolated from blood samples and one of the main causes of infections in intensive-care units [3]. It is a commensal organism existing in low densities in the intestinal environment of more than 50% of all humans. Most infections with *C. albicans* are caused endogenously but contagion from human to human is also possible. Examples for diseases caused by this fungus are infections of skin and nails or the mucous membrane of mouth, oesophagus and vagina. It can also be responsible for invasive infections like pneumonia, blood poisoning or meningitis.
A. fumigatus is a ubiquitous mould. In contrast to *C. albicans* it is highly thermotolerant (the fungus can live in up to 70°C) and its conidia (spores) survive long periods without water. Nearly all infections with this fungus are caused by inhalation of conidia. A contagion from human to human occurs only rarely. The conidia are gray-green and 2.5 - 3 μm in diameter, so they can easily reach the lung alveoli. *A. fumigatus* is the germ of three types of diseases: saprophytic aspergillosis (aspergillome), allergic aspergillosis and invasive aspergillosis. Aspergillomes are ”balls” of mycelia in lung cavities of patients that formerly suffered from tuberculosis, pneumonia or other lung diseases. Allergic aspergillosis shows symptoms like asthma and fibrotic changes of lung tissue which result in a reduced lung functionality. Invasive aspergillosis is a very dangerous disease for immunosuppressed patients with a mortality rate of up to 90%. It mostly starts in the lung but the fungus can spread and is able to infect other organs like heart, liver, kidneys or brain. A recent publication [4] reviews knowledge about *A. fumigatus* and factors contributing to its virulence. More information on the fungus can also be found at [5].

There are lots of databases and data warehouse systems in many different disciplines of research, especially in the fields of biology, molecular biology, medicine and of course bioinformatics. In 2007 one can find about 968 different databases in the internet that are important for biological research [6], a number which is growing fast. Some of them are well known, for example those of the NCBI (National Center of Biotechnology, [7]) like GenBank or PubMed. They are knowledge based, which means they store biological knowledge derived from different sources and they cover many organisms. In contrast, the databases ArrayExpress [8] or GEO (Gene Expression Omnibus, [9]) are storing experimental data. There are also databases that gain access to very special information on only few organisms, organs or other sample types. An example is EyeSite, which contains protein families of the eye. For the work with pathogenic fungi the databases CADRE [10] and e-fungi [11] are especially important. Both are maintained by the University of Manchester and collaborators. CADRE was built to store and analyse the genome of *A. fumigatus* and subsequently of other *Aspergillus* subspecies like *A. nidulans*. E-fungi stores genome and functional genomics data like protein interaction data and metabolic pathways of several fungal species like *A. fumigatus* and *C. albicans*.

We established a new data warehouse for experimental data from different cellular levels of human-pathogenic fungi. It stores datasets created by different working groups. This allows other groups to reanalyse those data under new points of view. It also enables comparisons between data from transcriptome and proteome. These integrated analyses will facilitate new insights into internal processes of the fungi and reveal details that are relevant for fungal infection. This will eventually lead to improved prevention, diagnosis and therapy of fungal diseases.

2 Methods

2.1 Schema

The new data warehouse at the Hans-Knoell-Institute (HKI) is based on Protecs (Decodon, [12]). Protecs itself is based on the postrelational database system Caché by Intersystems. Postrelational means that data are not only stored in tables but also in an object model. So they can be accessed via SQL queries but also via their object ID, property or method. This allows fast transactions and an easy development of user interfaces and application programmes. The data warehouse can be accessed through a web interface and several Java-applications. Its concept is shown in Figure 1. Genomic data of fungal pathogens are stored together with experimental transcriptomic and proteomic data and some microbiological information. Analyses can be applied on them so new knowledge can be gained.

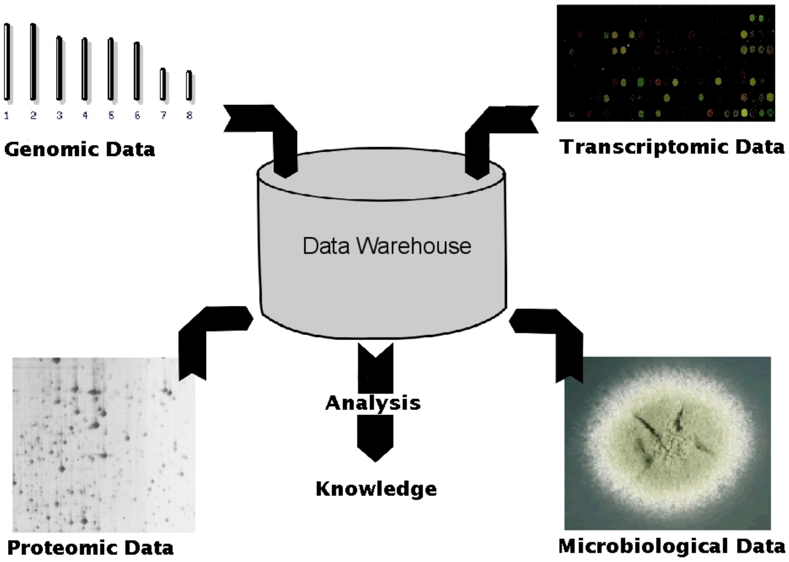


Figure 1: Concept of the data warehouse

Up to now data from three organisms, namely *Aspergillus fumigatus*, *Aspergillus nidulans* and *Candida albicans*, are stored in the database. *A. fumigatus* and *C. albicans* are the most important fungal pathogens for humans and *A. nidulans* is the model organism for filamentous fungi. All three species are main research topics of the Hans-Knoell-Institute. Hence, data are easily available and appropriate to establish and test the functionalities of the data warehouse. It is planned to include other pathogenic fungi as well as data from human as fungal host in the future.

2.2 Implementation

The entire datasets in the data warehouse are based upon the NCBI data content. They contain all the details provided by GenBank [13] like gene names, nucleotide sequence, protein product, GO annotation [14], etc. The genomic data are inserted into the data warehouse by a separate tool which only the administrator of the database is permitted to use. This is important to maintain the consistency of all basic data without any uncertainties.

This tool is also used for updates. Data in GenBank are updated regularly, for example when new names or synonyms for genes are available. These modifications have to be reinserted into the database. The import tool automatically updates only those data where it is necessary. These changes are carried out on all database entries so that experimental data are still cross-linked through the current genomic data.

All further data are built up hierarchically on that basis (Figure 2). The currently investigated organism is always on top of the hierarchy. Before inserting experimental data, a project has to be created. Within the project the user has to define some cultivations including descriptions. Each cultivation can consist of one or several samples. For every sample one or more microarrays or 2D gels can be defined and several scans of the same image can be included. The hierarchy models the experimental structures very well.

Data from different cellular levels can be combined while creating this hierarchy. The data warehouse provides the possibility to insert array and gel data obtained from the same sample. This experimental setup is optimal to achieve highly comparable results. It is also possible to work with data from a certain cellular level and to obtain corresponding data of other cellular levels from other working groups. Then data will be represented as two different cultivations.

In summary, the hierarchical structure allows keeping the overview over all inserted data. In addition, it helps scientists to understand the experiments of others. Hence, it minimises the amount of necessary coordination and supports well documented work flows.

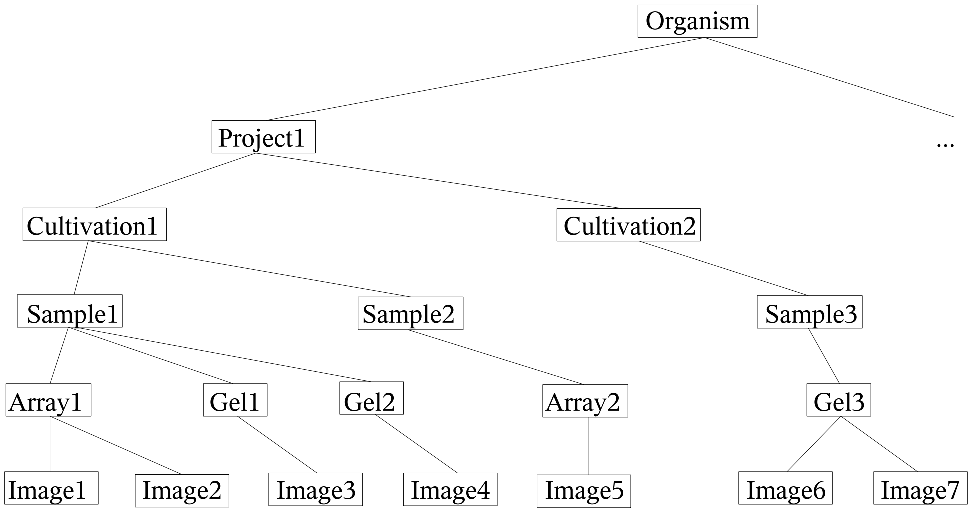


Figure 2: Hierarchical structure of the data warehouse

When inserting the different steps of the hierarchy into the database, the user has the possibility to define numerous parameters (Table 1). Therefore, users can store information about every experimental step in great detail. This accuracy again aids the efficient reuse of data by other scientists. Names and meanings of the parameters were chosen according to the Microarray Gene Expression Data Ontology (MO, [15]). The MGED Society defines standards for working with microarrays. For proteomic data such a widely accepted standard is still missing. The Proteomics Standards Initiative (PSI) is developing standards for several parts of proteomics experiments. They work closely together with MGED. In the future there might be a standard for storing data from both cellular levels. This standard will be adopted by our data warehouse as soon as it is published. At least on the first three hierarchical steps, parameters are very similar for transcriptomic and proteomic data. So MO can be applied to proteomic data in this context as well.

Some of the parameters can only contain numerical values, are strings or represent dates. Others are implemented as string lists. These lists represent the most common experimental conditions and were created according to the needs of microbiologists. This method assures that the most important values are included. It also helps to keep the database entries homogeneous. This is a further step to make reanalyses of data more efficient for other researchers. For further projects new parameters and new list values can be added. Additionally, on every hierarchical level information about the generation of a certain piece of data is stored. Author, date and time are recorded, so every user can see by whom and when a sample was made or a gel was run.

Table 1: Parameters of the top three steps of the hierarchy

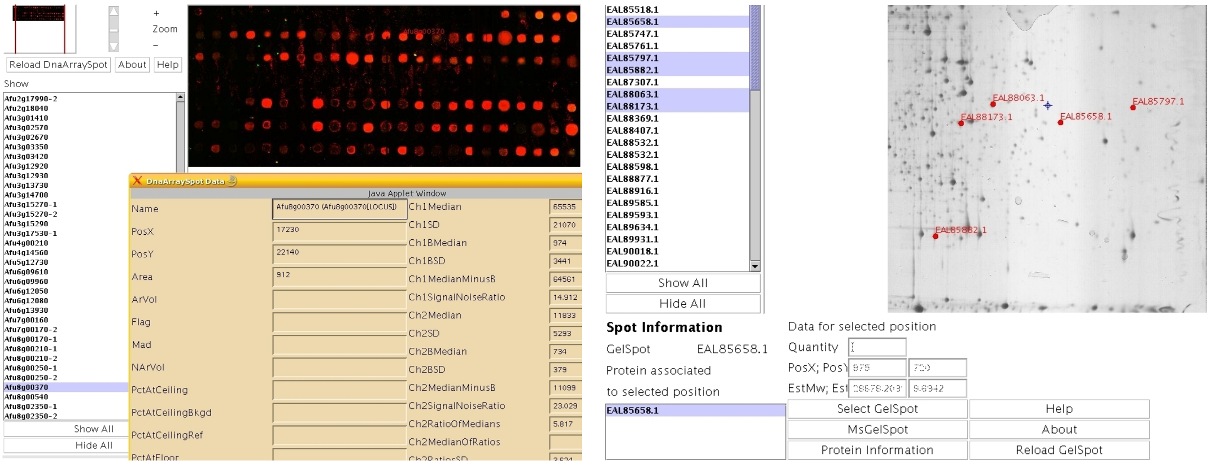
Project	**Cultivation**	**Sample**
Date_ended (date)	Incubate (str)	Amount of applied protein (μg) (num)
Date_started (date)	IndividualGeneticCharacteristics (str)	Fractionate (str list)
Goal (str)	MediumType (str list)	LabelCompound (str list)
Sponsor (str)	Number of with MS identified spots (num)	LabID (str)
	Perturbation (str)	LysisBuffer (str list)
	SubstrateType (str list)	Purify (str list)
		Temperature (°C) (num)
		TimePoint (min) (num)

All parameters of one hierarchical level are collected in parameter sheets. Several sheets can be organised into a topic. This topic can be selected immediately after logging into the database. So a user does not have to see all parameters. One rather can simply choose the topic appropriate to an experimental setup. The division of parameters into different topics will become especially important when first human data are inserted into the database. Experiments made with samples from humans are very different from those made with fungi. Hence, lots of new parameters will emerge. For this reason, the data warehouse structure was built to be highly adaptable.

2.3 Visualisation

One very important functionality of the data warehouse is the visualisation of data. This works in two different ways. First, a user can search for a special gene or protein of an organism and view the arrays or gels on which it was detected. The search functionality can be used by entering a name, synonym name or ID of an external database for a gene or protein. All these information have to be inserted into the data warehouse via GenBank files or manual annotation beforehand. In this way, one can directly compare the different values of a gene on different arrays or view time-courses in different arrays or gels of a project. This is especially suited for researchers interested in one particular gene or protein who look for additional information to own experiments. Another way is to browse within the hierarchy of a project and view a special array or gel. If part of the array or gel name is known, direct searching is possible again. Else the user has to select the project level to find the right project. Subsequently, he has to choose a cultivation and then a sample. For the selected sample all corresponding arrays and gels are

listed. Their images can be inspected and data can be compared to other samples. This is very helpful when one is interested in the overall behaviour of several genes or proteins at one special experimental setup or in comparison to other experimental setups.
The result of searches are array or gel images which display the regulation of genes or proteins of one sample under special experimental conditions.



Afu2g17990-2
Afu2g18040
Afu3g01410
Afu3g02570
Afu3g02670
Afu3g03350
Afu3g03420
Afu3g12920
Afu3g12930
Afu3g13730
Afu3g14700
Afu3g15270-1
Afu3g15270-2
Afu3g15290
Afu3g17530-1
Afu4g00210
Afu4g14560
Afu5g12730
Afu6g09610
Afu6g09960
Afu6g12050
Afu6g12080
Afu6g13930
Afu7g00160
Afu7g00170-2
Afu8g00170-1
Afu8g00210-1
Afu8g00210-2
Afu8g00250-1
Afu8g00250-2
Afu8g00370
Afu8g00540
Afu8g02350-1
Afu8g02350-2

Name	Afu8g00370 (Afu8g00370[LOCUS])	Ch1Median	65535
PosX	17230	Ch1SD	21070
PosY	22140	Ch1BMedian	974
Area	912	Ch1BSD	3441
ArVol		Ch1MedianMinusB	64561
Flag		Ch1SignalNoiseRatio	14.912
Mad		Ch2Median	11833
NArVol		Ch2SD	5293
PctAtCeiling		Ch2BMedian	734
PctAtCeilingBkgd		Ch2BSD	379
PctAtCeilingRef		Ch2MedianMinusB	11099
PctAtFloor		Ch2SignalNoiseRatio	23.029
		Ch2RatioOfMedians	5.817
		Ch2MedianOfRatios	
		Ch2RatiosSD	3.624

EAL85518.1
EAL85658.1
EAL85747.1
EAL85761.1
EAL85797.1
EAL85882.1
EAL87307.1
EAL88063.1
EAL88173.1
EAL88369.1
EAL88407.1
EAL88532.1
EAL88532.1
EAL88598.1
EAL88877.1
EAL88916.1
EAL89585.1
EAL89593.1
EAL89634.1
EAL89931.1
EAL90018.1
EAL90022.1

Spot Information

GelSpot EAL85658.1

Protein associated to selected position

EAL85658.1

Data for selected position

Quantity

PosX; PosY 975 720

EstMw; Est 28678.203 9.6942

(a) (b)

Figure 3: Clickable view of a microarray (a) and a 2D gel (b)

By clicking on an image, spot information are displayed as lists and can easily be used for analyses. All information imported from several image analysis programmes can be seen. Figure 3a shows array data from GenePix [16]. All data that are usable for further analyses are stored and displayed in a table at click on an array image. Figure 3b displays the view on 2D gel data. Again by clicking on a gel spot the user gets information about protein identifications or post-translational modifications. Also mass spectra, peak lists and hit lists can be stored allowing to trace protein identifications back to the raw data.
To make further information on genes and proteins available, the data warehouse stores accession numbers of external databases. It can translate between different accession number systems and even between several versions of one external database. This allows to view spots with current names of proteins that were annotated with other names several months ago. That way, the data warehouse displays all experiments where a certain protein was found independent of the timeliness of the name used for searching.

Proteomic and transcriptomic data can be compared because they are cross-linked within the data warehouse. If in an experiment a special protein shows significant changes in its abundance, the user can investigate, whether this protein shows interesting regulation in other experiments of this project, too. One can also find out whether this protein plays an important role within one of the other projects. Additionally, there is a link from the protein to its gene. Here, the user can check whether the according mRNA had been spotted on any array within the data warehouse and if there are any interesting changes in its intensities. This is especially important when transcriptomic and proteomic experiments have been made using the same organism and the same experimental conditions. Thus, direct comparisons between transcriptomic and proteomic data are possible.

2.4 Security

The management of users and their access rights is very important for every database. When a user inserts some data, he has to define access rights to every other user concerning his data. To simplify the handling of access rights, users can be grouped into roles. Access rights can be assigned to those roles instead of defining them for every single user. This way it is possible to define who is allowed to read data and who should be permitted to edit them. Every part of a project can be made available to different subsets of registered users or can be kept completely private. The user can decide which data of a project can be viewed by the public, which data should be available to collaborators and which should be visible only to himself. This way, one can easily protect data and assure that nobody changes data by mistake. This is especially important when the database will be made fully available via the internet because nobody wants ongoing work to be seen by everyone and an internet user should have no more than read-only privileges.

2.5 Import of experimental data

The last step of the inserting process after defining experimental structure and access rights is the import of the data themselves. The data warehouse is meant for the storage of experimental data, like scanned images and raw quantitation data analysed by image analysis software. Array and gel images can be imported from all major file formats, including TIFF and JPG. So it nearly does not matter by which scanner and in which format an image was produced. Images of 2D gels can be viewed as clickable web images. The image can be downloaded in its original form as well as a tiled image which is advantageous for image processing software. Array images of a single-channel array are treated like gel images. For dual-channel images it is necessary to upload the image of every channel separately. The data warehouse calculates a composed image where both channels are overlaid by using an internal processing method specially adapted to such images. This overlaid image can be viewed as web image and can be downloaded like the original single images.

The import of corresponding quantitation data is more complex. Data formats of six different image analysis programs can be imported at present. These are tab-delimited text files formatted according to GenePix Array List (GAL), formats of ArrayVision (GAL and SG), as well as Affymetrix CDF for microarray data. Exported data of 2D gel image analysis software such as Delta2D as well as XML exports of DeCyder and ImageMaster 2D Platinum can be inserted for proteomic data. Image analysis with Delta2D can be tightly integrated with the data warehouse: all data such as analysis setups, spot quantities, and warpings can directly be transferred into the database. The data warehouse will be extended to support other data formats in the future.

2.6 Standardised export

The export functionality of the data warehouse is also extremely important for the reanalysis of data. All experimental data that have been uploaded once can be downloaded again in appropriate formats.

The export of array data is made according to Minimum Information About a Microarray Experiment (MIAME, [17]) guidelines using MicroArray Gene Expression - Markup Language (MAGE-ML, [18]). This microarray data exchange format was chosen as a standard for gene expression and is supported by the MGED. It is build up according to the MicroArray Gene Expression - Object Model (MAGE-OM). This is a model for data exchange that has been formulated using the Unified Modelling Language (UML). UML is the industry-standard, object oriented modelling language. MAGE-ML has been implemented using eXtensible Markup Language (XML). MGED ontology is closely linked to MIAME and enables unmistakably interpretation of microarray annotations, as well as searches in microarray databases. Some journals today require that papers reporting microarray experiments must be accompanied by MIAME. Data have to be be stored in a MIAME accepting database such as ArrayExpress as standard part of the publication process.

The export of 2D gel data at present is carried out as specified within the Proteomics Experiment Data Repository schema (PEDRo, [19]). This schema is a starting point for the development of a standardised data format similar to MAGE-ML. It provides possibilities to store proteomic data and information about their creation in a standardised manner. The according data interchange format is the Proteomics Experiment Markup Language (PEML). It is formulated using UML and implemented in XML with eXtensible Stylesheet Language Transformations (XSLT). XSLT transforms PEML files to be readable as HyperText Markup Language (HTML) by standard web browsers. It also translates old-format PEML files into new-format ones when the schema is modified. PEDRO is not further developed now, it will be replaced by PSI standards in the near future. We will switch to these standards as soon as they are established.

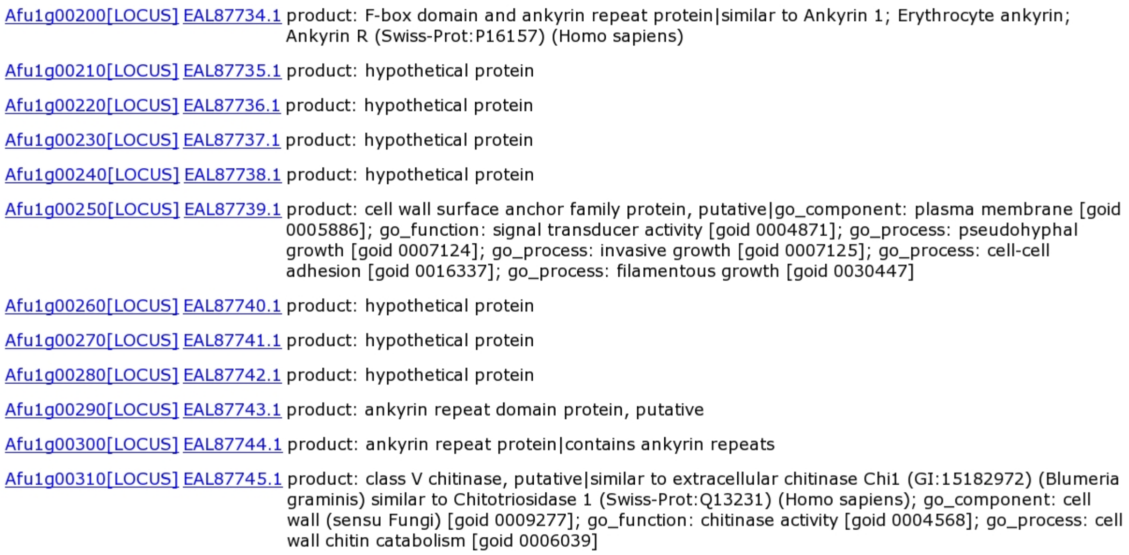
Via these two models, data formats are standardised and can then be imported into other databases or into appropriate analysis programmes like R [20] for further investigation. This facilitates the communication between this data warehouse and external tools or even other data warehouses. It allows other researchers to make their own analyses of existing data without the time and money consuming process of repeating the experiments.

3 Application

Time series data of a temperature-shift experiment on *Aspergillus fumigatus* were analysed in an integrated way for the first time. In this experiment liquid cultures of the fungus cultivated at 30°C were shifted to 48°C. The aim was to identify genes and proteins that play a role in the thermotolerance of the fungus. Additionally, we wanted to find similarities and differences between transcriptome and proteome level of the cells.

The transcriptomic data have been described previously [21] and were obtained from ArrayExpress [22] in a preprocessed form. They have already been subject of a reverse engineering approach to reconstruct the underlying genetic network [23]. The proteomic data were produced by Kniemeyer et al. using an optimised protocol [24] and are still unpublished. Gel images were analysed by the software package DeCyder [25] and raw data from this software were further processed and analysed using R. Proteins and transcripts were regarded as differentially expressed when they satisfied the following criterion: The logarithmised ratio of the normalised intensities of two different time points in the time course must have a higher absolute value than 90% of all intensities. That means that this ratio must be outside of the range of $[m-1.645*sd; m+1.645*sd]$ of all intensities, where m is the mean and sd is the standard deviation. These values can be computed by using Z-ratios [26]. Doing so, only a p-value of 0.1 is achieved. Because the analyses are still ongoing we kept this value to get a broader overview over the cellular processes. The exact procedure of creating and analysing the proteomic data and their comparison with the transcriptomic data will be reported elsewhere.

Proteomic and transcriptomic data were connected via the data warehouse. Both types of data were imported. The data warehouse provides the possibility to get gene names, protein names and functional information within a table (Figure 4). So one can find differentially regulated proteins and corresponding genes in one step. Then the user can make comparisons between regulation on both cellular levels. With the additional information about the putative function of proteins including some GO data, the user can get insights into the cellular processes that were affected in an experiment. A rough clustering of interesting genes and proteins into different categories is also possible.

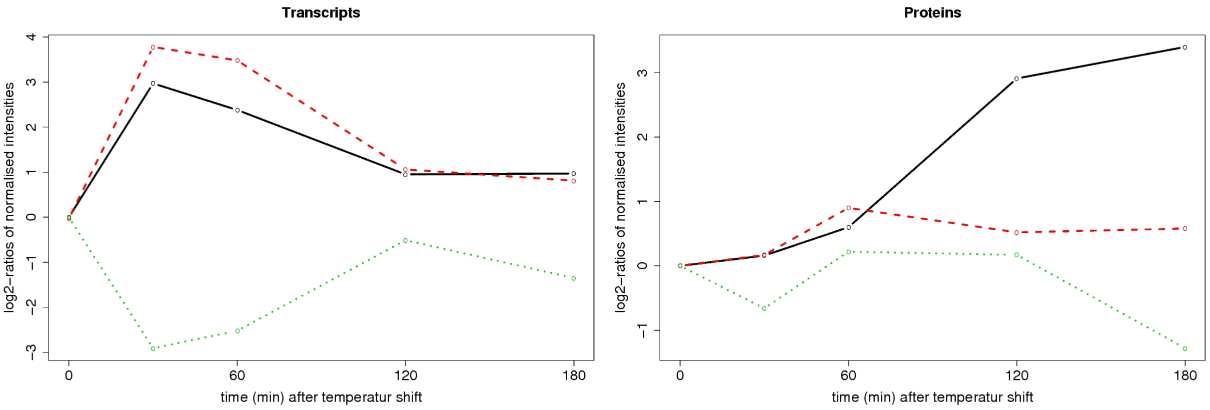


Afu1g00200[LOCUS]	EAL87734.1	product: F-box domain and ankyrin repeat protein|similar to Ankyrin 1; Erythrocyte ankyrin; Ankyrin R (Swiss-Prot:P16157) (Homo sapiens)
Afu1g00210[LOCUS]	EAL87735.1	product: hypothetical protein
Afu1g00220[LOCUS]	EAL87736.1	product: hypothetical protein
Afu1g00230[LOCUS]	EAL87737.1	product: hypothetical protein
Afu1g00240[LOCUS]	EAL87738.1	product: hypothetical protein
Afu1g00250[LOCUS]	EAL87739.1	product: cell wall surface anchor family protein, putative|go_component: plasma membrane [goid 0005886]; go_function: signal transducer activity [goid 0004871]; go_process: pseudohyphal growth [goid 0007124]; go_process: invasive growth [goid 0007125]; go_process: cell-cell adhesion [goid 0016337]; go_process: filamentous growth [goid 0030447]
Afu1g00260[LOCUS]	EAL87740.1	product: hypothetical protein
Afu1g00270[LOCUS]	EAL87741.1	product: hypothetical protein
Afu1g00280[LOCUS]	EAL87742.1	product: hypothetical protein
Afu1g00290[LOCUS]	EAL87743.1	product: ankyrin repeat domain protein, putative
Afu1g00300[LOCUS]	EAL87744.1	product: ankyrin repeat protein|contains ankyrin repeats
Afu1g00310[LOCUS]	EAL87745.1	product: class V chitinase, putative|similar to extracellular chitinase Chi1 (GI:15182972) (Blumeria graminis) similar to Chitotriosidase 1 (Swiss-Prot:Q13231) (Homo sapiens); go_component: cell wall (sensu Fungi) [goid 0009277]; go_function: chitinase activity [goid 0004568]; go_process: cell wall chitin catabolism [goid 0006039]

Figure 4: Connection between genes and proteins via the data warehouse; gene name = Afu..., [LOCUS] meaning that there is no actual gene name in GenBank so it was extracted out of the locus tag; protein name = EAL...; putative function = product:...

In this case study transcripts and proteins of the temperature shift experiment were analysed separately. Differentially expressed proteins and corresponding genes were found using the data warehouse. Subsequently, their time courses were compared.

We found 16 differentially regulated proteins for which the transcripts also showed differential expression. Many of them are heat shock proteins and chaperones. As it was expected, they are clearly upregulated on both cellular levels (e.g., 30 kDA heat shock protein, Figure 5, solid). Other upregulated proteins belong to several metabolic pathways, like the pyruvate metabolism or are regulators of the cell cycle. Here, the upregulation on transcriptomic level is still clearly visible, whereas on proteomic level it is less clear, but still apparent (e.g., myo-inositol-phosphate synthase, putative, dashed in Figure 5). Four proteins and their corresponding transcripts were found to be downregulated (e.g., 5- methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase, pointed in Figure 5). They are involved in biosyntheses or assembly of ribosomes, respectively.
In addition, 23 differentially regulated proteins were found for which the transcripts did not show remarkable changes. About 580 transcripts were found for which the associated proteins did not show such changes or could not be detected on any gel.



(a) (b)

Figure 5: Time courses of selected transcripts (a) and proteins (b); black (solid) = 30 kDA heat shock protein; red (dashed) = myo-inositol-phosphate synthase, putative; green (pointed) = 5-methyltetrahydropteroyltriglutamate–homocysteine S-methyltransferase

Figure 5 also shows that the maximal differential expression on proteomic level is delayed in comparison to the transcriptomic level. This is plausible, because transcription occurs before translation. So transcripts have to accumulate first before their protein products can be synthesised.

4 Conclusions

This publication shows that an integrated analysis of time series data from transcriptome and proteome level is necessary to get new insights into fungal pathogenesis. This comprehensive view can only be achieved by using both types of data. They clearly complement each other and provide additional information. It is shown that a data warehouse can be helpful in such an integrative approach by providing the possibility to store those data and to visualise them using cross-links.

5 Discussion

For some differentially regulated proteins the corresponding transcripts were differentially expressed, too. But for some more differentially regulated proteins the transcripts did not show differentiall regulation and vice versa. So by using only data from one cellular level, researchers might miss interesting genes or proteins that could shed new light on a scientific question. In the future, when more datasets will be included into the data warehouse, it will also be interesting to compare genes and proteins of this temperature shift experiment with other experimental conditions. This will help to learn more about their functions and regulation. Hence, a data warehouse for the storage of a diverse set of experimental data assures that all interesting facts are accessible and knowledge can be extracted.

In many cases transcripts showed the same tendency in regulation as their proteins (i.e., up or down), but the time courses often showed big differences (Figure 5). This implies that in further investigations the time courses have to be analysed in detail to find the direct connection between transcriptome and proteome.

Also the integration of data from both cellular levels is very simple at present. Data are analysed separately and after finding proteins which seem to be differentially expressed the corresponding genes are examined and then the regulation is compared. First attempts were made to bring both types of data together by using biplots [27] (data not shown). This is part of a collaboration with the Fraunhofer Institute for Interfacial Engineering and Biotechnology in Stuttgart and the German Cancer Research Center (DKFZ) in Heidelberg. The data warehousing system MCHiPS [28] in Heidelberg is able to store transcriptomic and proteomic data and applies Correspondence Analysis (CA, [29]) on them. We will establish a permanent connection from our data warehouse to MCHiPS, so all users can take advantage of the CA tool.

All analyses to select differentially expressed genes and proteins have to be made outside of the database environment at the moment. A user has to export data from the data warehouse to analyse it by some external tools. It is planned to include such tools into the data warehouse to advance the visualisation and analysis of the data within the database. This should be possible in an easy and comfortable way because of the object oriented access provided by Caché. Mostly R scripts will be used because it is open source and provides a Java interface which can be used to connect to the database.

At present only microarray data, 2D gel data and mass spectrometry data can be inserted into the data warehouse. In the future it is planned to support gel free proteomics like LC-MALDI-MS and LC-ESI-MS/MS. These and related approaches to analyse proteomic data can be an alternative to the classical 2D gels for some special research topics. Also other kinds of data may play an important role in investigations of human-pathogenic fungi and will have to be included into the data warehouse in the future. Examples are fungal growth kinetics or microscopic images of nucleophiles or macrophages interacting with fungal spores or hyphae.

A data warehouse is most beneficial when many researchers are using it. Then lots of data are available through it. So this data warehouse will be opened to collaborators first and later also to the public. This includes cooperations with other groups and the connection of some other data warehouses to this one. Important partners are two data warehouses for fungi maintained in Manchester, CADRE and e-fungi. Another collaboration is the work within Eurofungbase [30]. This concerted action wants to build up and maintain an integrated, durable European genomic database required for innovative research on filamentous fungi. This database will become a centre for related systems and could be integrated and preserved in a centralised European genomic database. For all collaborations the importing and exporting functionality of the data warehouse will be improved so that the exchange of data is made easy and as complete as possible.

6 Acknowledgements

Part of this work was supported by the German Research Foundation DFG within the Priority Programme 1160.

The authors would like to thank the reviewers for their comments and advises to improve this manuscript.

References

[1] M. M. McNeil et al. Trends in mortality due to invasive mycotic diseases in the United States, 1980-1997. *Clin Infect Dis*, 33 (5), 641-647, 2001.

[2] J. Hacker (Ed.). Menschen, Seuchen und Mikroben: Infektionen und ihre Erreger. 1st edition, *Beck, Munich*, 2003.

[3] H. Hahn, D. Falke, S. H. E. Kaufmann (Eds.). Medizinische Mikrobiologie und Infektiologie. 5th edition, *Springer, Heidelberg*, 2005.

[4] A. A. Brakhage. Systemic fungal infections caused by *Aspergillus* species: epidemiology, infection process and virulence determinants. *Curr Drug Targets*, 6 (8), 875-886, 2005

[5] The *Aspergillus* Website, `http://www.aspergillus.org.uk/`

[6] M. Y. Galperin. The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res*, 35 (Database issue), D3-D4, 2007.

[7] D. L. Wheeler et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 34 (Database issue) ,D173-D180, 2006. `http://www.ncbi.nih.gov`

[8] H. Parkinson et al. ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 33 (Database issue), D553-D555, 2005. `http://www.ebi.ac.uk/arrayexpress`

[9] T Barrett et al. NCBI GEO: mining millions of expression profiles–database and tools.*Nucleic Acids Res*, 33 (Database issue), D562-D566,2005 `http://www.ncbi.nlm.nih.gov/geo`

[10] J. E. Mabey et al. CADRE: the Central *Aspergillus* Data REpository. *Nucleic Acids Res*, 32 (Database issue), D401-D405, 2004. `http://www.cadre.man.ac.uk`

[11] M. Cornell et al. e-Fungi: An e-Science Infrastructure for Comparative Functional Genomics in Fungal Species. *AHM2005*, 2005. `http://www.e-fungi.org.uk`

[12] Decodon GmbH, Greifswald. The Protecs brochure. `http://www.decodon.com/Solutions/Protecs`

[13] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler DL. GenBank. *Nucleic Acids Res*, 34 (Database issue), D16-D20, 2006.

[14] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet* 25, 25-29, 2000. `http://www.geneontology.org`

[15] P. L. Whetzel. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22 (7), 866-873, 2006.
`http://mged.sourceforge.net/ontologies/index.php`

[16] Molecular Devices. GenePix Pro 6.0 Manual. *Molecular Devices*, 2006

[17] A. Brazma et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4), 365-371, 2001.
`http://www.mged.org/Workgroups/MIAME/miame.html`

[18] P. T. Spellman et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9), 2002.
`http://www.mged.org/Workgroups/MAGE/mage.html`

[19] C. F. Taylor CF et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotech*, 21, 247-254, 2003.
`http://pedro.cs.manchester.ac.uk/`

[20] I. Ross, G. Robert. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 3, 299-314, 1996.
The R Project for Statistical Computing. `http://www.r-project.org/`

[21] W. C. Nierman et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, 438 (7071), 1151-1156, 2005.

[22] H.S. Kim et al. ArrayExpress, accession number E-MEXP-332. 2005

[23] R. Guthke, O. Kniemeyer, D. Albrecht, A. A. Brakhage, U. Moeller. Discovery of Gene Regulatory Networks in *Aspergillus fumigatus*. *Lecture Notes in Bioinformatics*, 4366, 22-41, 2007.

[24] O. Kniemeyer, F. Lessing, O. Scheibner, C. Hertweck, A. A. Brakhage. Optimisation of a 2-D gel electrophoresis protocol for the human-pathogenic fungus *Aspergillus fumigatus*. *Curr. Genet*, 49 (3), 178-189, 2005.

[25] Amersham Biosciences. DeCyder Differential Analysis Software, Version 5.0, User Manual, *Amersham Biosciences, Sweden*, 2003.

[26] C. Cheadle, M. P. Vawter, W. J. Freed, K. G. Becker. Analysis of microarray data using Z score transformation. *J Mol Diagn*, 5(2), 73-81, 2003.

[27] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58 (3), 453-467, 1971.

[28] K. Fellenberg, N. C. Hauser, B. Brors, J. D. Hoheisel, M. Vingron. Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. *Bioinformatics*, 18, 423-433, 2002. `http://www.mchips.org`

[29] K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel, M. Vingron. Correspondence anaylsis applied to microarray data.*Proc Natl Acad Sci USA*, 98: 10781-10786, 2001

[30] Eurofungbase, `http://www.eurofung.net`