An integrated platform for analyzing molecular-biological data within clinical studies

Toralf Kirsten[1], Jörg Lange[1], Erhard Rahm[1,2]

{tkirsten,lange}@izbi.uni-leipzig.de,
rahm@informatik.uni-leipzig.de

[1] Interdisciplinary Center for Bioinformatics Leipzig, University of Leipzig
[2] Dept. of Computer Science, University of Leipzig

Abstract. To investigate molecular-biological causes and effects of diseases and their therapies it becomes increasingly important to combine data from clinical trials with high volumes of experimental genetic data and annotations. We present our approach to integrate such data for two large collaborative cancer research studies in Germany. Our platform interconnects a commercial study management system (eRN) with a data warehouse-based gene expression analysis system (GeWare). We utilize a generic approach to import different anonymized pathological and patient-related annotations into the warehouse. The platform also integrates different forms of experimental data and public molecular-biological annotation data and thus supports a wide range of genetic analyses for both clinical and non-clinical parameters.

1 Introduction

Clinical trials help to study the cure process and survival rate of patients for new or modified therapies and drugs, e.g. to deal with specific types of cancer. For this purpose, many patient and treatment parameters are observed and analyzed. In addition to analyzing the success of entire therapies, one can also find parameters acting as classifiers, for which participating patients show a different therapy course and success. On the other hand, diseases and therapy processes are deeply affected by molecular-biological conditions for genes, proteins and their complex inter- and intracellular interactions. For instance, cancer cells underlie genomic mutations and thus have a modified gene expression that is often increased in higher states of the disease. To better understand the genotype-phenotype interrelationships for diseases and their therapies it becomes increasingly important to combine clinical and molecular-biological data, e.g. to investigate the relationship between pathological classifications and genomic disparities [Co03]. These studies utilize new experimental high throughput techniques for patients like microarray-based gene expression analysis [Ka05]. An ultimate goal is to support personalized therapies with respect to individual genetic patient conditions.

The need to combine clinical and molecular-biological data poses specific data integration requirements. So far these different types of data are not only maintained in

a variety of different data sources but are also managed by different complex data management and analysis systems. Clinical trials typically involve many institutions and complex workflows. They are usually managed by commercial study management software, such as eResearch Network[1] (eRN), Oracle Clinical[2], and MACRO[3]. Most of these systems are certified by public authorities, such as *Federal Drug Administration* (FDA) in the USA and *European Medicines Agency* (EMEA) in Europe [Ku03]. On the other hand, molecular-biological experimental data is typically maintained in specific genomic databases, such as ArrayExpress [Bra03], Stanford Microarray Database (SMD) [She01], and Gene Expression Omnibus (GEO) [Ba05]. They support the analysis of huge amounts of gene expression data but without considering clinical parameters. In addition, there are numerous publicly available data sources providing annotations for molecular-biological analysis, e.g. Entrez [Ma05], SwissProt [Ba04], GeneOntology [GOC04], and OMIM [OMIM00].

Overviews of currently available approaches and tools for data integration in bioinformatics are given in [St03, LC03]. Most of the approaches focus on the integration of publicly available annotation data. [Na04] proposes a data warehouse platform to integrate patient-related data with data from different types of molecular-biological experiments and annotations. However, the platform is limited in the number of annotation sources and does not support clinical trials across different institutions. NCICB (National Cancer Institute Center for Bioinformatics) has started a large biomedical data integration effort within the caBIG initiative *(*cancer Biomedical Informatics Grid) [Bu05,Co03].

In this paper, we present our analysis platform integrating clinical and molecular-biological data for two large collaborative cancer research studies in Germany. One study aims at investigating molecular mechanisms of malignant lymphoma[4], the other focuses on glioma[5]. First results [Hu06] are recently published. Our platform interconnects the commercial study management system eRN with a data warehouse-based gene expression analysis platform (GeWare). We utilize a generic approach to import different pathological and anonymized patient-related annotations into the warehouse where it is used for improved data analysis. The platform also supports integration of different forms of experimental data and public molecular-biological annotation data. We believe our approach is quite general and applicable in similar research studies on analyzing molecular mechanisms for different types of diseases and therapies.

In the next section we introduce the project environment and resulting requirements. Section 3 presents the overall architecture of our integration approach and platform. In section 4 we present our generic approach to import and maintain annotations. Section 5 explains the multidimensional data warehouse model and different analysis capabilities before we conclude.

---

[1] http://www.ert.com

[2] http://www.oracle.com/industries/life_sciences/clinical.html

[3] http://www.infermed.com/macro/

[4] http://www.lymphome.de/en/Projects/MMML/index.jsp

[5] http://www.gliomnetzwerk.de/

2 Project Requirements

Clinical trials typically involve complex workflows across different organizations. Figure 1 visualizes some process portion of a clinical trial focusing on major data acquisition steps. It starts with the identification of relevant patients to participate in the clinical trial based on defined inclusion criteria. These criteria have to be carefully specified to select patients relevant for the respective research question while preserving enough patients to support statistically valid analysis. For selected patients *personal data* is captured, such as age, sex, material status or non-/smoker distinction. Some properties reflect habits and peculiarities of patients that can have a great impact in the later analysis, e.g. when the data is partitioned in non-/smoker portions.

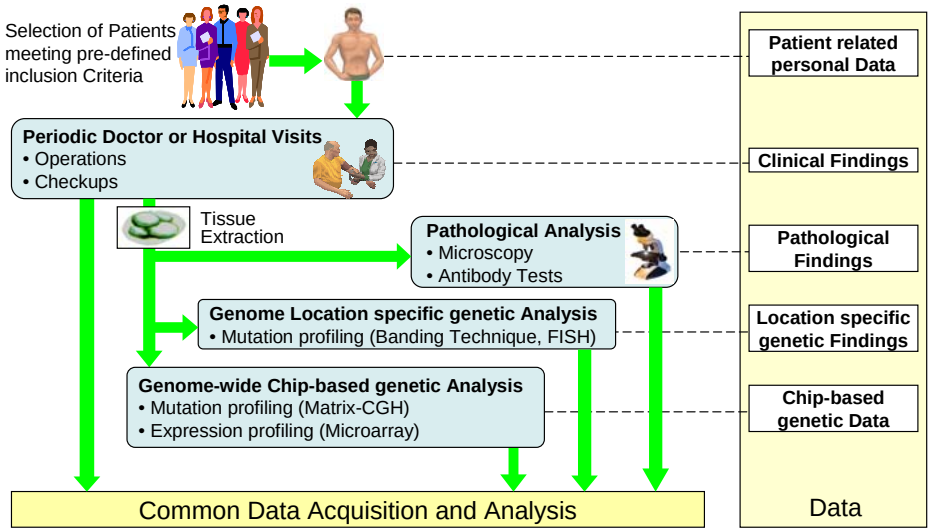

Figure 1: Project environment and resulting data

A *clinical finding* is produced whenever a patient visits a doctor or the hospital. That can happen regularly, e.g. for quarterly checkups, or when an adverse event happens. In both cases, the clinical finding describes the current clinical state of the patient and makes it possible to track the therapy status by utilizing precisely defined parameters. Typically, such clinical findings are stored in a study management system. In addition, it could be necessary to extract diseased tissue material for a patient within an operation, e.g. cancer nodes. This material is then analyzed by pathologists, e.g. using light microscopy or antibody tests. The pathologists describe the properties of the extracted tissue material and hence create a *pathological finding* that can influence the decisions of doctors in the therapy process.

Moreover, parts of the extracted tissue material can be utilized to experimentally measure properties at the genetic level, particularly using expression profiling and mutation profiling. *Expression profiling* studies the so-called expression behavior (activity) of interesting genes w.r.t. different conditions, e.g. healthy vs. diseased

tissues or for different points in time. Microarrays [She95, Lo96] are the currently prevalent tools measuring the expression of thousands of genes at the same time.

The second experimental approach, mutation profiling, focuses on the genetic diversity of patients. Normally, genes are located at fixed positions on a chromosome. However, individual mutations (insertions, deletions, moves) of sequences can have a significant impact on the development and therapy of diseases. This holds particularly for large block-wise mutations, such as copies and movements across different chromosomes. Current techniques to measure such genetic imbalances include the banding analysis [Ca70], the Fluorescent in situ-Hybridization (FISH) [Me95], and Matrix-based comparative genomic hybridization (Matrix-CGH) [Ka92]. The first two techniques focus on a specific genome location and bring out a relative small number of data or just a description. By contrast, the Microarray-based gene expression and the Matrix-CGH mutation profiling operate genome-wide and, hence, generate huge amounts of data. Typically, the banding and FISH analyses are performed in different hospitals, while the Microarray-based expression and Matrix-CGH mutation profiling are centrally conducted by specialized labs.

Requirements

The sketched project environment and workflow require a comprehensive and standardized approach to integrate the different types of data and to perform data analysis. The specific requirements are:

- **Data integration**: The different kinds of data obtained from the described clinical workflow need to be integrated for analysis, in particular personal data, several types of findings, and molecular-biological data produced by high-throughput techniques. The high volume of experimental data asks for a central management of the integrated data. To enhance the analysis capabilities it is also desirable to integrate molecular-biological annotation data from publicly available sources.
- **Utilization of existing information systems**: Typically, commercial study management systems are utilized to manage patient-related personal data and her corresponding finding data, whereas different genomic databases manage expression and mutation profiling data. In order to save time and cost such already existing systems should be used and connected instead of designing a new comprehensive system from scratch.
- **Uniform data specification**: Data of different steps such as clinical and pathological findings are generated in different hospitals and organizations. To keep the data comparable it is imperative to enforce uniform data acquisition procedures and standardized data formats. This concerns not only the metadata such as the sets of parameters to be provided but also the permissible data (instance) values and their meaning. The latter may be enforced by conformed vocabularies.
- **Autonomous data input**: Manual data input into paper forms should largely be avoided and replaced by direct data entry into the study management system. The data entry should be autonomously take place where the data is generated by using pre-defined web templates. The study management can cen-

trally store the data and should perform extensive validity tests to ensure high data quality.

- **Central molecular-biological experiments**: Molecular biological experiments should be performed in a central laboratory for each type of experimental data. This ensures uniform laboratory conditions and device properties as needed for a comparative analysis of experimental data.
- **Privacy aspects**: Legal requirements demand the protection of the patients' privacy. In particular, identifying data such as id card number, social insurance number or the person's name must not be stored together with other data, e.g. clinical and pathological findings.
- **Advanced data analysis**: Comprehensive molecular-biological data analysis should be supported for different theoretical and biological researchers to fully leverage the collected and integrated set of data.

3 Platform Architecture

To meet these requirements for two large collaborative cancer research studies we have developed a comprehensive data integration and analysis platform at the University of Leipzig. Figure 2 shows the overall architecture of this platform. It interconnects two existing data management systems, the study management system eRN and the gene expression warehouse GeWare [KHR04]. Both systems themselves integrate data from several sources, permit interactive user input and support analysis of their data.

The study management system eRN allows users at participating institutions to autonomously specify patient-related personal, clinical, and pathological data using predefined web forms. To enforce the anonymity of patient-related data, a *technical patient identifier* is generated whenever a new patient enters the clinical study. All personal identifications such as patient names or social security number are excluded and only anonymous patient data tagged with the technical patient identifier is entered in the study management system. To support high data quality the system implements different rule-based input and consistency checks (e.g. minimum and maximum values) as well as cross validations. Specific data validation reports indicate input imbalances or missing data to be corrected by users before the data is accepted and made available for analysis. All analysis routines on study management data can be performed via web interfaces but are typically restricted to basic statistical reports (e.g., number of examined patients at various stages of the therapy).

While the eRN system manages patient-related data, the GeWare system deals with chip-based expression and mutation data. Currently, this data is generated at central labs by Microarray-based and Matrix-CGH chip experiments. This data is much more voluminous than the patient-related data and cannot be stored within eRN. GeWare provides web interfaces to upload new experimental data and to specify their technical annotations on laboratory conditions, such as hybridization temperature.

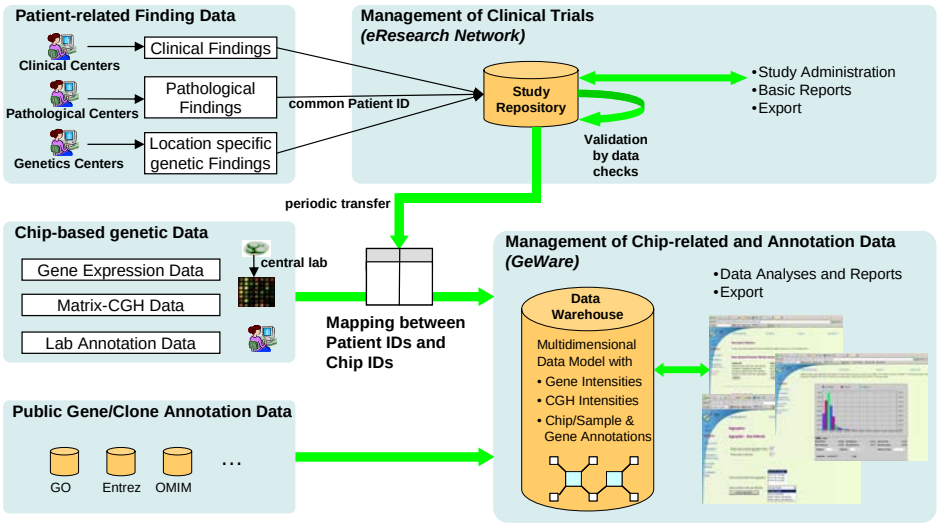

Figure 2: Overall Architecture of the Platform

To combine patient-related data with chip-based data for combined analysis, GeWare also imports a subset of patient-related data from eRN. The selection depends on the research project and currently subsumes about 100 to 130 parameter values per patient. While the patient-related data is identified by the patient identifier, the chip-based data utilizes a chip identifier from which the patient identifier can not be derived. We thus provide a *mapping table* associating each chip identifier with the corresponding patient identifier to correctly combine clinical, pathological and experimental data and to permit an over-spanning data analysis. In addition, GeWare integrates publicly available gene/clone annotation data for extended analysis possibilities. This data integration is performed by a query mediator approach and outlined in [Ki05].

GeWare comprises different reports and analysis methods. For instance, it is possible to find lists of differentially expressed genes according to different clinical circumstances by analyzing experimentally generated data together with biological and selected patient-related annotations. Furthermore, data can be exported for external analysis by specialized statistical or data mining software.

The platform not only preserves the anonymity of personal data but also utilizes a sophisticated authentication and authorization concept for different user groups. In particular, access rights can be granted/revoked not only for the access to both systems and its data (patient-related annotations and experimental data), but also for the functions on the data, such as import, export, query etc. According to the user profile, e.g. doctors in a hospital, pathologists and biostatisticians, the web user interface is automatically generated to only cover the allowed functions of both systems.

4 Annotation Integration

Depending on the clinical focus, the trials can be conducted and documented in different ways. For instance, clinical lymphoma studies usually describe diseased cancer nodes using parameters such as node size and node type, but also the state of thyroids due to its important role in the metabolism. On the other hand, in glioma studies the specific brain region is important to annotate. Hence, the captured parameters typically differ between studies. Similarly, annotations of experimental conditions for microarray data and Matrix-CGH arrays may differ substantially. While standards like MIAME[6]/MIAME-CGH [Bra01] give a recommendation about the minimal information to be captured, they do not specify what values should be used for each parameter. Hence, additional standardizations are needed to avoid non-comparable or conflicting data.

Addressing these problems thus requires support for different sets of annotations for different studies and consistent data values. For this purpose, our platform provides a *generic* approach to specify and maintain annotations so that adding or changing annotation specifications are easily possible. For these specifications we provide so-called *annotation templates* to prescribe the parameters to be annotated and controlled vocabularies to constrain permissible parameter values. A template consists of pages that group together related parameters, e.g. for personal data, pathological findings or experimental parameters. Each page can be hierarchically organized. Annotation parameters and their corresponding values (metadata and data) are stored generically using the so-called Generic Annotation Model (GAM) introduced in [DR04]. These approaches for specifying and storing annotations avoid changing the database schema for new or changed annotations. This makes it easy to support additional clinical studies or additional types of annotations in our platform.

Figure 3 illustrates the process to specify annotations for a clinical study, to map annotation data from the study management system eRN to GeWare, and to use (query) annotations in GeWare. Initially, the annotation parameters for which values have to be captured in a new clinical trial are specified (Step 1). Furthermore, the study management system eRN is configured to manage data for the clinical trial (Step 2a). In addition, the subset of parameters to be transferred from eRN to GeWare for analysis purposes are specified (Step 2b). Based on these parameters a new template can then be created in GeWare (Step 3) consisting of hierarchically arranged pages. Based on the database schema of eRN and the tree-based annotation schema of the template, a schema mapping (Step 4) is created for the relevant subset of parameters. The result of this schema mapping associates each source element, i.e. the parameter specific attribute and table of the relational database schema of eRN, with the corresponding target element, i.e. the parameter-specific path in the annotation schema. While this schema mapping is currently performed manually, it could also be done semi-automatically by utilizing schema matching algorithms [RB01]. The resulting schema mapping is regularly used to transfer new patient-related annotation values from eRN to GeWare (Step 5).

[6] MIAME stands for Minimal Information about a Microarray Experiment.

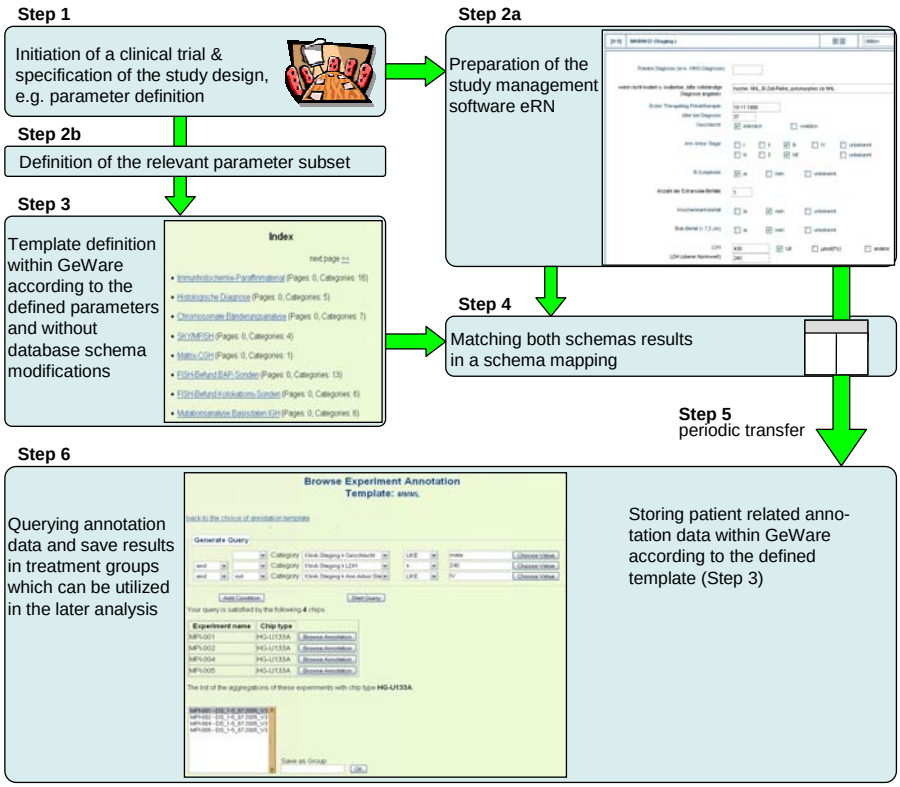

Figure 3: Defining, transferring, and querying patient-related annotations

GeWare allows browsing through the annotations, querying them and applying them to extract and analyze experimental data (Step 6). For querying, the user can define multiple conditions that are combined with the logical operators AND, OR, and NOT. The query result identifies lists of chips, patients or genes that can be used to specify experimental data portions for further analysis.

5 Multidimensional data model and analysis

GeWare is a relational data warehouse integrating and maintaining both annotation data and experimental data of different types. Figure 4 shows a high-level view of its multidimensional schema built of dimension and fact tables. Experimental data like numerical expression values are stored in fact tables containing the majority of data. Dimensions provide information on the meaning of facts and are needed for their analysis. In particular they maintain selected annotation data on genes and chips / patients. Multidimensional modeling is a proven approach for data warehouses supporting flexible and fast analysis for large data volumes.

Our schema contains separate fact tables for genes and clones and for transformed and analyzed data values. The *Gene Intensity* fact table stores transformed gene expression data, whereas the *Clone Intensity* fact table contains intensity data of determined clones. Additional fact tables are kept for *Expression* and *CGH Matrices*, to store the intensities of those genes (clones) participating in gene (clone) groups determined by a specific analysis method, such as clustering.

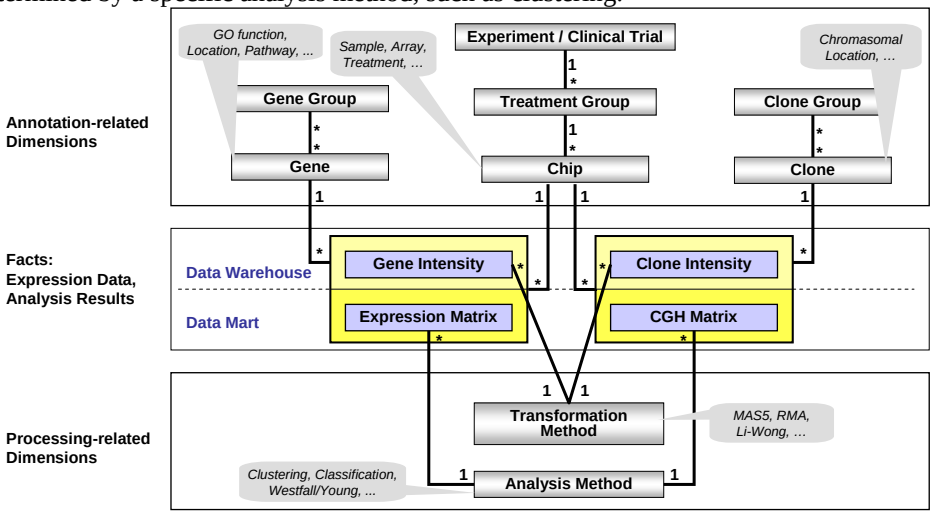

Figure 4: Multidimensional data warehouse model

The dimensions can be grouped into annotation- and processing-related dimensions, which are shown in Figure 4 together with some illustrating examples. Annotation-related dimensions include tables for g*enes*, *chips*, and *clones* (and thus patients) and their groupings. Processing-related dimensions specify *transformation* and *analysis methods* describing the computational methods and their parameters used to compute gene/clone intensities, and to determine gene (clone) groups for expression (CGH) matrices, respectively.

Dimensions can be organized into generalization/specialization hierarchies to provide different levels of abstraction for analysis. For example, the chip dimension is organized into three levels, experiment, treatment group and chip. Experiment is the most abstract level describing a clinical trial encompassing many chips which can be grouped into so-called treatment groups. Each of these treatment groups may include chips for a specific experimental condition, e.g. for tissue probes from the same patient at a specific time point in the clinical trial.

The sketched multidimensional data model supports high analysis flexibility. While current approaches typically evaluate a complete data matrix, i.e. gene expression and CGH matrix, containing the intensity values for all measured genes/clones and several/all chips, we now can focus on individual or comparative analysis to an arbitrary subset of intensity values determined by specific annotation values of interest. The selection may be based on a value at a specific level of a single dimension or any combination for several dimensions, e.g. to consider both gene and clinical parameters. A variety of analysis methods can be applied to the selected data set, e.g.

queries to determine differentially expressed genes or clustering to find co-expressed genes. Such result groups can in turn be saved for further queries and analysis. The platform also supports the export of the pre-computed analysis results, e.g. gene/clone groups and expression/CGH matrices, to perform analysis in external tools.

As an example for a combined analysis of experimental and annotation data, Figure 5 shows a gene expression heatmap for a selected group of 25 genes (rows) and a treatment group of 25 chips/patients (columns). Furthermore, the expression data is analyzed by hierarchical clustering for both, chips and genes. The dendrogram on the top represents the chip hierarchy while the one on the left hand side shows the gene hierarchy. In addition, a classification of the chip data by pre-defined classifiers, in this case the cancer stage which was acquired by the clinical diagnoses, using available patient-related annotation data is visualized by a colored band (different colors represent different values) above the heatmap. Thus the user can determine if there is a correlation between the hierarchical order resulting from the clustering and the fragmentation stemming from the classification.

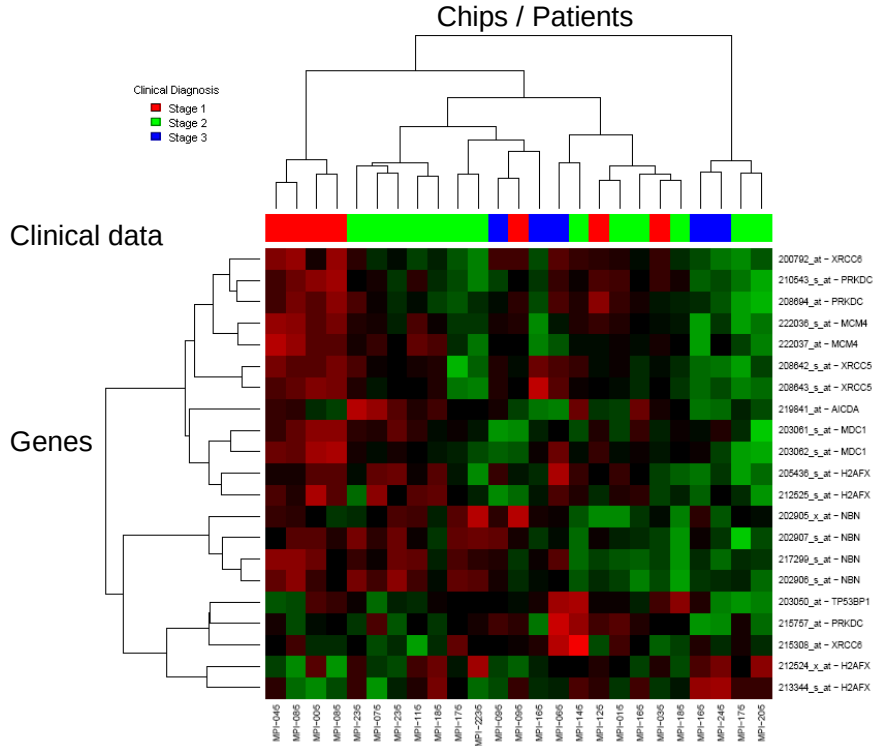

Figure 5: Heatmap utilizing patient-related annotations

6 Conclusions

We presented a platform combining clinical and molecular-biological data for large-scale collaborative clinical research studies. The approach combines two proven subsystems for managing clinical trials and gene expression analysis. The clinical study system uniformly captures patient-related data from several participating hospitals. All patient-related data is kept in anonymous form and is interrelated with other data by a technical patient id only. The warehouse-based platform imports selected clinical annotations from the study system and combines them with data of centrally performed molecular-biological high-throughput experiments. Annotations are managed generically to easily support different studies and changing analysis needs. Currently, the platform is fully operational and is in use in two large clinical collaborative research projects in Germany.

Acknowledgements

The authors are thankful for useful hints and discussions with Markus Löffler and Hilmar Berger who are involved in the clinical studies and made this work possible. We would also like to thank Hans Binder for fruitful discussions. The work is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) grant BIZ 1/3-1 and the German Cancer Aid (Deutsche Krebshilfe) grant 70-3173-Tr3.

References

[Ba04] Bairoch, A. et al.: Swiss-Prot: Juggling between evolution and stability. Briefings in Bioinformatics. 5:39-55, 2004

[Ba05] Barrett, Tanya et al.: NCBI GEO: mining millions of expression profiles - database and tools. Nucleic Acids Research, 33: D562-D566 (Database issue), 2005

[Bra01] Brazma, Alvis et al.: Minimum Information about a Microarray Experiment (MIAME) – Towards Standards for Microarray Data. Nature Genetics, 19, 2001

[Bra03] Brazma, Alvis et al.: ArrayExpress: A public database of gene expression data at EBI. C. R. Biologies, 326(10-11), 1075-8

[Bu05] Buetov, Kenneth H.: Cyberinfrastructure: Empowering a Third Way in Biomedical Research. Science, 308: 821-24, 2005

[Ca70] Caspersson et al., Identification of human chromosomes by DNA-binding fluorescent agents, Chromosoma 30: 215 - 227, 1970

[Co03] Covitz, P.A.: Class struggle: Expression profiling and categorizing cancer. The Pharmacogenomcs Journal, 3:257-60, 2003

[DR04] Do, Hong-Hai; Rahm, Erhard: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Proc. EDBT, Heraklion, Greece, Springer LNCS, March 2004.

[GOC04] The Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 32: D258-D261, 2004

[Hu06] Hummel, M. et al.: Transcriptional and genomic profiling provides a biological definition of Burkitt lymphoma and identifies novel prognostic groups within mature aggressive B-cell lymphoma. (submitted)

[Ka92] Kallioniemi, A. et al.: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science, 258(5083): 818-21, 1992

[Ka05] Kallioniemi, O: Dissection of molecular pathways of cancer by high-throughput biochip technologies and RNA interference. Breast Cancer Research, 7:43, 2005

[KHR04] Kirsten, T., Do, H.-H.; Rahm, E.: A Data Warehouse for Gene Expression Analysis. Technical Report. Univ. of Leipzig, 2004

[Ki05] Kirsten, T.; Do, H.-H.; Rahm, E.; Körner, C.: Hybrid Integration of molecular biological Annotation Data. Proc. 2nd Int. Workshop on Data Integration in the Life Sciences, San Diego, 2005

[Ku03] Kuchinke, W.; Ohmann C., „eTrials“ werden zur Routine, Deutsches Ärzteblatt 2003; 100:A 3081-3084, 2003

[LC03] Lacroix, Zoe; Critchlow, Terence: Bioinformatics: Managing Scientific Data. Morgan Kaufmann, 2003

[Lo96] Lockhart, D.J. et al: Expression Monitoring by Hybridization to High-density Oligonucleotide Arrays. Nature Biotechnology 14, 1996

[Ma05] Maglott, Donna et al.: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research, 33: D54-D58, (Database issue), 2005

[Me95] Mecucci, C.: FISH (fluorescent in situ hybridization): the second youth of cytogenetics. Haematologica, 80(2):95-7, 1995

[Na04] Nagarajan, R; Ahmed, Mushtaq; Phatak, Aditya: Database Challenges in the Integration of Biomedical Data Sets. Proc. of the 30th VLDB Conf., Toronto, Canada, 2004

[OMIM00] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000

[RB01] Rahm, Erhard; Bernstein, Philip A.: A survey of approaches to automatic schema matching. VLDB Journal, 10(4):334-50, 2001

[She95] Shena, M. et al: Quantitative Monitoring of Gene Expression Patterns with a complementary DNA Microarray. Science 270, 1995

[She01] Sherlock, Gavin et al.: The Stanford Microarray Database. Nucleic Acid Research, 29(1), 2001

[St03] Stein, L.: Integrating Biological Databases. Nature Review Genetics, 4(5): 337-45, 2003