An integrated analysis platform for experimental and clinical data in modern cancer research studies

Lange J[1], Kirsten T[1], Rahm E[2]
[1]*Interdisciplinary Centre for Bioinformatics, University of Leipzig, Germany*
[2]*Departement of Computer Science, University of Leipzig, Germany*
lange@izbi.uni-leipzig.de

Introduction To investigate molecular-genetic causes and effects of diseases and their therapies it becomes increasingly important to combine data from clinical trials with high volumes of experimental data generated using various chip technologies and their annotations. We present our approach to integrate such data for two large collaborative cancer research studies in Germany – the Molecular Mechanism of Malignant Lymphoma (MMML) and the German Glioma Network. Our platform interconnects a commercial study management system (eRN) with a data warehouse-based gene expression analysis system (GeWare) [1]. 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-genetic annotation data and thus supports a wide range of collaborative analyses for both clinical and non-clinical parameters.

Methods We have developed a comprehensive data integration and analysis platform at the University of Leipzig interconnecting two existing data management systems. On the one hand the study management system eRN allows users at participating institutions to remotely enter all data typically handled in traditional clinical trials e.g. patient-related personal, clinical, and pathological data. To support high data quality the system implements different rule-based input and consistency checks which indicate input imbalances or missing data to be corrected by users. On the other hand the GeWare system deals with chip-based gene expression and array-CGH data and comprises different reports and analysis methods. Chip 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 further annotations like laboratory parameters. To combine patient-related data with chip-based data for combined analysis, GeWare also imports a subset of patient-related data from eRN in a generic manner using so called annotation templates. 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 [2].

Results We established a warehouse-based platform combining clinical experimental chip data for large-scale collaborative cancer research studies and based on two dedicated subsystems for managing clinical trials and gene expression analysis. Selected clinical annotations were imported by daily transfer from the study system and combined with data of centrally performed molecular-biological high-throughput experiments. Annotations are managed generically to easily support different studies and changing analysis needs. Grouping functions for genes, probes and samples that can be used later within analyses are available. Interactive Analyses for data visualization (e.g. heatmaps as displayed in Fig.1) allow a quick overview for hypothesis generation and statistic reports indicate significant values of the large-scale array data. Furthermore desired data can be extracted for specific analyses outside the platform.

Discussion The analysis platform described here proved to be a valuable tool for storing, accessing and analysing high-dimensional gene expression and array-CGH data together with clinical, histopathological and other experimental data. The web-based interface allows interactive analyses for experimenters and the results are stored for further methods. The platform runs successfully within the cancer project MMML and will be extended for the aims of the German Glioma Network

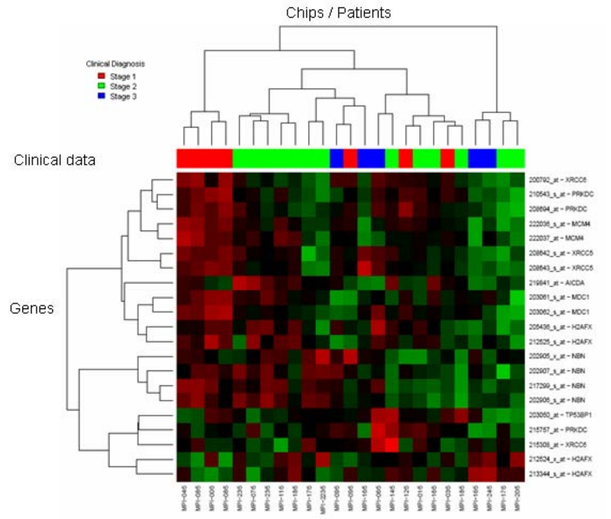

Fig. 1: This figure shows a gene expression heatmap for a selected group of 25 genes (rows) and a treatment group of 25 chips/patients (columns). Furthermore, the expression data is analyzed by hierarchical clustering for both, chips and genes. The dendrogram on the top represents the chip hierarchy 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 in the study, is visualized by a colored band above the heatmap. Thus the user can determine here if there is a correlation between the hierarchical order resulting from the clustering and the fragmentation stemming from the classification.

Literature

[1] Kirsten T, Lange J, Rahm E. An integrated platform for analyzing molecular-biological data within clinical studies. Proc. Workshop Information Integration in Healthcare Applications at 10th EDBT Conference, Munich. March 2006

[2] 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 Science, San Diego, Springer LNBI, 2005