Integrating clinical, gene expression, protein expression and preanalytical data for in silico cancer research.

Delphine Rossille, Anita Burgun, Céline Pangault-Lorho, Thierry Fest

Integrating clinical, gene expression, protein expression and preanalytical data for *in silico* cancer research

Delphine ROSSILLE [a,c,1], Anita BURGUN [a,c], Céline PANGAULT-LORHO [b] and Thierry FEST [b]

[a] *UPRES EA3888 Conceptual Modeling of Biomedical Knowledge, Rennes University, France*

[b] *INSERM U917 Microenvironnement & Cancer, HITC Department, Rennes Hospital, France*

[c] *Medical Information Department,Rennes Hospital, France*

Abstract : We present the phase I development of an integrative platform for the analysis of clinical, gene expression, protein expression and pre-analytical data. The platform is aimed at providing transparent access and analysis tools to researchers investigating new biomarkers and prognosis factors in the particular field of lymphoma diseases. In this article, we report on the data integration phase. The platform's principal advantage is its completeness as it integrates in a single environment clinical, genomic and proteomic data, allowing for their combined analysis. The architecture consists in a data warehouse including data on patients, clinical trials and array platforms and a DeMilitarized Zone for data exchange. A secure web-based platform allows any collaborative team to request the data warehouse and access basic statistics on integrated data. The presented system is currently in use.

Keywords : Systems integration; Information storage and retrieval; Databases, Genetic; Cancer

Introduction

Nowadays, an increasing amount of genomic and annotation data is produced, stored worldwide and analyzed for better understanding of diseases and investigating new biomarkers. After independently studying genomic, proteomic or clinical data it has become clear that simultaneous analysis, by increasing the analysis power, is the next step. However biologists are faced with the problem of exploiting such massive heterogeneous and distributed data. They need systems to transparently access relevant data and to facilitate their interpretation thanks to appropriate analysis tools.

The BMS-Ly national research project, conducted by Rennes hospital, is devoted to investigating blood biomarkers and prognosis factors of the Diffuse Large B-Cell non-Hodgkin Lymphoma (DLBCL) disease, more specifically predictors of answers to treatments. The study is based on data collected at the diagnosis, that is before patients received any treatment. Our contribution to this project aims at supporting biostatisticians by setting up an integrative platform for the analysis of all the produced

[1] Corresponding Author : Delphine ROSSILLE, Département d'Information Médicale, CHU Rennes, 2 rue Henri Le Guilloux, 35033 Rennes, France, Email : delphine.rossille@chu-rennes.fr

data (clinical, gene and protein expression data and preanalytical parameters). We report here on the phase I development involving the data integration. Phase II will be concerned with the development of tailored analysis methods, based on statistics and data mining.

1. Background

Bioinformatics requires the development of informatics systems to facilitate the handling, querying and exploration of high throughput data. The largest project is probably CaCore [1] that provides data standards and system infrastructures for cancer research. Since the explosion of high-throughput technologies, many tailored data management systems have been developed. Some are deposit and retrieval systems for published well-annotated experimental data like [2] for gene expression data of any species and pathology or [3] for gene and protein expression data. Others are specific to a pathology ([4] for clinical and gene expression data of solid cancers) or a medical domain ([5] for gene and protein expression data for immune cells). Few integrate gene and protein expression data [3, 6]. Some public data repository and retrieval systems provide in addition online analysis tools (Array Express with the Expression Profiler online tool [7], ITTACA [8] with bioinformatics analysis for clinical and gene expression data). Only few laboratory databases include clinical, gene and protein expression data in their systems for complex analysis ([9], [10]).

2. Material and Method

2.1.Material

The cohort will consist in 300 patients and 150 healthy control subjects. The patients' dataset is collected from the GOELAMS (Groupe Ouest-Est des Leucémies aigües et Autres Maladies du Sang) 02-03 and 075 clinical trials, two French nationwide multicenter studies launched in 2005. The local medical doctor asks the GOELAMS for the inclusion of a new patient. Having checked the inclusion criteria, the GOELAMS trial manager provides an anonymized inclusion number to be used from then on. The local medical doctor accesses the electronic Case Report Form (eCRF) online system including built-in password security. The collected data include clinical, demographic, biological data, received treatments, abnormal events or randomization information and are stored in a database in Nancy. The blood samples collected at the inclusion stage are processed in Rennes before being sent for transcriptomic or proteomic analysis.

Preanalytical data, including the delays between the sampling and freezing processes, are collected by Rennes to check their impact on the samples quality [11]. The transcriptomic data are produced by Montpellier Affymetrix microarray platform and sent to Rennes in their raw proprietary format (.cell files). They are derived from Affymetrix Human Exon 1.0 ST arrays Hu-Ex-1_0-st, arrays providing the most comprehensive coverage of the genome, including not only well annotated regions. The protein expression data are produced by Grenoble Bio-Rad SELDI-TOF Mass Spectrometry platform. Produced spectra, detected peaks and experimental conditions are sent to Rennes. All experimental data are related to their patients thanks to an anonymization number.

2.2.Method

As the patients' data are distributed in different data sources, an appropriate system's architecture was determined in accordance with required functionalities. The database conceptual modeling was developed so as not to be restricted to clinical trials, the investigators wishing for future use to store data from patients having the same pathology but not included in the clinical trial. The platform was then implemented.

3. Results

3.1.Required Functionalities

The study involves five French research teams that will actively collaborate. Four teams produce the clinical, gene and protein expression and pre-analytical data. Two teams will investigate the data. The data will be analyzed with commercial software (on raw expression data files) or with tailored analysis methods (based on statistical and data mining methods), hence data must be available in a unique format. Finally the Rennes hospital requested all data to be banked for future research.

3.2.System's Architecture

We chose to integrate all the data into a tailored data warehouse, for centralizing into a unique system with a unified data schema. This approach, compared to federated or mediation systems, ensures consistency among data, rapid queries on massive amount of data and fast complex analyses. The main drawback is when data updates are required. However in our case, only exceptional updates could be needed and all transactions are historically recorded in a log book. Data elements are integrated in the data warehouse and files of raw experimental data and experimental conditions are stored in a data server and linked to the anonymized inclusion numbers of their patients. The data warehouse includes array platforms and clinical trial information.

The platform is a secure web site accessible by all the teams (Figure 1). A DMZ zone is used to download by secure FTPs the produced data files, for their latter integration into the database. Gene annotations are accessible by hyperlinks to public databases from the probeset identifier for transcriptomic data.

Platform: PhP, HTML , javascript - Data warehouse: Oracle 9i relational database.

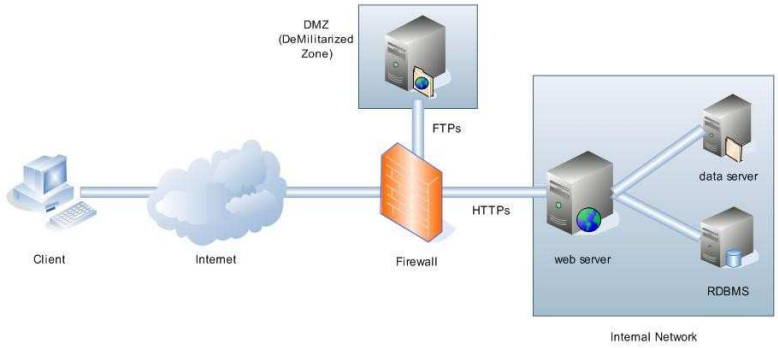

Figure1: System's architecture

3.3.Conceptual Data Modeling

The model (Figure 2) consists in six data spaces: four patient-related (clinical, gene expression, protein expression and pre-analytical data spaces), the clinical trials and the array platforms spaces. In the future other spaces (publications, gene annotations or biomarker results) could easily become part of the model. The modeling is generic enough to integrate patients not included in clinical trials.

Hence a patient is integrated for a medical *episode* related to a specific *disease*, an *episode* consists in several *stages*. At each stage the patient can have *treatments*, *exams results*, pre-analytical and expression data or *events* such as status aggravation. For instance, in the case of the GOELAMS trial, <*typeEpisode*> is 'clinical trial' and the first <*typeStage*> is 'inclusion' to which are associated pre-analytical parameters, gene and protein expression data and the clinical data collected at the inclusion stage. Pre-analytical data are associated to expression data as it informs on the sample quality (Table *PreAnalyticalParameters*). Each *TranscData* is related to a probeset identifier (attribute *id_probe*), and an *ArrayPlatform* consists in a list of probeset identifiers as labelled by the manufacturer (table *probesets*), and general information on its design name (attribute *nameArray*), manufacturer, version and species. The raw files of each genomic or proteomic analysis are linked to its experimental conditions (for instance attribute *transcRawData_fileName* in table *TranscExperimentalConditions*).

3.4.The Platform's functionalities

- All data, except for proteomic data (not yet implemented), are automatically integrated into the database from formatted files. Clinical data must be integrated first in order to create the patients. The gene expression data can only be integrated if the corresponding array platform is registered into the database.
- Descriptive statistics of the database status are displayed according to criteria s.a. sex or age, allowing the analyser to check data integration over time.
- Data for patients' groups can be exported according to clinical criteria. Hence combined analysis of expression and clinical data is made possible. The corresponding list of raw data files is produced and allows their analysis through commercial software.
- Authentication and access rights : In addition to data anonymization, the secure web site and the DMZ are accessible through Rennes hospital's firewall with login and password authentification (Figure 1). Access rights to the web site are either for "user" or "administrator", the last one allowing to integrate data into the database, register new array platforms, reschedule integration of past files and check the log book.

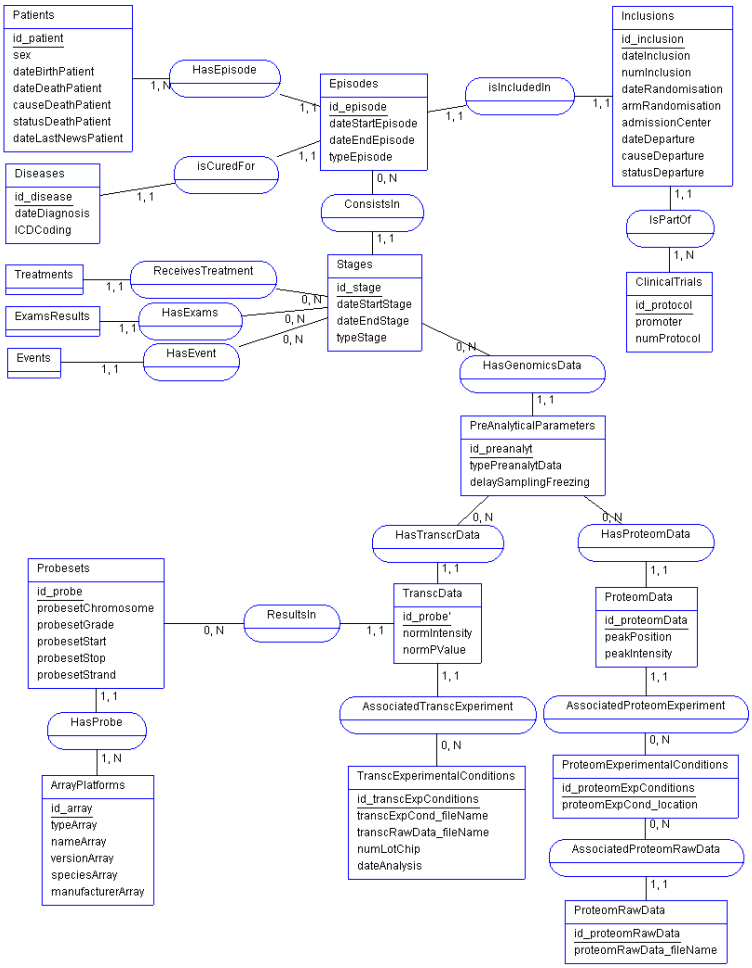

Figure 2: Data Modeling

4. Discussion and Conclusion

The national French BMS-Ly research project was funded to discover biomarkers from blood samples of patients diagnosed with DLBCL. We present the data integration platform built to support researchers by making data access transparent regardless of their heterogeneity. As of today, the presented platform is in use, the integration of protein expression data remaining to be implemented. Security of patients' data has been taken care of: anonymized inclusion numbers, firewall and secure architecture. Phase II will focus on the development of tailored analysis tools.

In regards of the literature on similar platforms [2-8], by integrating pre-analytical, clinical, genomic and proteomic data in a unified data model, the platform's main advantage is its data completeness. Indeed the majority of published platforms store gene or protein expression data, to the exception of [3,5-6]. Furthermore clinical data are usually limited in scope. Our application integrates all available clinical data from

the eCRF system. It includes information on the samples quality that can be taken into account during analyses. It integrates full exon expression data not all well-annotated or even not annotated at all, providing finer insights. The presented data warehouse is aimed not to be limited to its current application but will be most valuable in the future for simultaneous analysis of gene and protein expression data. Furthermore, contrary to public applications like [8] where genomic data are stored in files and only a description is kept in the database, the presented platform stores data as files and as data elements, making the focus on specific clinical and expression data elements simple. GeWare [12] is a similar integrative platform, with the difference of the data model used (multidimensional vs relational).

The presented platform can be improved. For instance, data on experimental conditions will be in the future compliant to standards, such as the MIAME standard for transcriptomic data. The platform has several limitations. Presently, the data to be integrated are well formatted: the clinical data are mostly based on controlled vocabularies specified by the GOELAMS, expression data are stored in proprietary formats. The syntax variations detected for some clinical data were taken care of. Semantic interoperability will however have to be overcome when patient's data will be derived from other sources. Ontologies like SNOMED CT, by offering the ability to relate concepts together, will be then most valuable. Finally the platform should accomodate cross-platform integration and analysis.

References

[1] P.A. Covitz *et al*, caCORE: A common infrastructure for cancer informatics, *Bioinformatics* **19** (18) (2003), 2404-2412

[2] H. Parkinson *et al*, ArrayExpress – a public repository for microarray gene expression data at the EBI, *Nucleic Acids Research* **33** (2005), D553-D555

[3] K. Ikeo *et al*, CIBEX: Center for Information Biology gene Expression database, *C.R.Biologies* **326** (2003), 1079-1082

[4] K. Kato *et al*, Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues, *Nucleic Acids Research* **33** (2005), D533-D536

[5] A. Hijikata *et al*, Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells, *Bioinformatics* **Sept. 25** (2007)

[6] C. Chelala *et al*, Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets. BMC Genomics 2007, 8:439

[7] M; Kapushesky *et al*, Expression Profiler: next generation – an online platform for analysis of microarray data, *Nucleic Acids Research* **32** (2004), W465-W470

[8] A. Elfilali *et al*, ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis, *Nucleic Acids Research* **34** (2006), D613-D616

[9] H. Hu *et al*, Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research, *Pharmacogenomics* **5** (7) (2004), 933-941

[10] A. Nagarajan *et al*, Database challenges in the integration of biomedical data sets, In *Proc. 30th VLDB Conference*, Toronto, Canada, 2004, 1202-1213

[11] C. Pangault *et al*, Stakes of pre-analytical parameters in blood transcriptomic and proteomic analysis: Application to clinical research: The GOELAMS trial, *Med Sci (Paris)* **3** (2007), Mar 23, 13-18(french)

[12] E. Rahm, The GeWare data warehouse platform for the analysis of molecular-biological and clinical data, *J. Integrative Bioinformatics*, **4** (1) (2007), journal.imbio.de