Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW

Émilie Guerin, G. Marquet, Anita Burgun, Olivier Loréal, Laure Berti-Équille, Ulf Leser, Fouzia Moussouni

Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW

E. Guérin [1], G. Marquet [2], A. Burgun [2], O.Loréal [1], L. Berti-Equille [3]

U. Leser [4], F. Moussouni [1]

[1] INSERM U522 CHU Pontchaillou, 35033 Rennes, France
[2] EA 3888 LIM, Faculté de Médecine 35043 Rennes, France
[3] IRISA, Campus Universitaire de Beaulieu, 35042 Rennes, France
[4] Dep. for Computer Science, Humboldt-Universität, 10099 Berlin Germany

Abstract. Researchers at the medical research institute Inserm U522[1], specialized in the liver, use high throughput technologies to diagnose liver disease states. They seek to identify the set of dysregulated genes in different physiopathological situations, along with the molecular regulation mechanisms involved in the occurrence of these diseases, leading at mid-term to new diagnostic and therapeutic tools. To be able to resolve such a complex question, one has to consider both data generated on the genes by in-house transcriptome experiments and annotations extracted from the many publicly available heterogeneous resources in Biomedicine. This paper presents GEDAW, a gene expression data warehouse that has been developed to assist such discovery processes. The distinctive feature of GEDAW is that it systematically integrates gene information from a multitude of structured data sources. Data sources include: i) XML records of GENBANK to annotate gene sequence features, integrated using a schema mapping approach, ii) an inhouse relational database that stores detailed experimental data on the liver genes and is a permanent source for providing expression levels to the warehouse without unnecessary details on the experiments, and iii) a semi-structured data source called BioMeKE-XML that provides for each gene its nomenclature, its functional annotation according to Gene Ontology, and its medical annotation according to the UMLS. Because GEDAW is a liver gene expression data warehouse, we have paid more attention to the medical knowledge to be able to correlate biology mechanisms and medical knowledge with experimental data. The paper discusses the data sources and the transformation process that is applied to resolve syntactic and semantic conflicts between the source format and the GEDAW schema.

1 Introduction

In human health and life science, the rapid emergence of new biotechnological platforms for high throughput investigations in genome, transcriptome and proteome, prompts further advances in information management techniques to take in charge the

[1] Regulation of functional balances of normal and pathological liver

data and knowledge generated by these technologies. A tremendous amount of bio-medical data is continuously deposited by scientists in public Web resources, and is in return searched by other scientists to interpret results and generate and test hypothesis.

The management of these data is challenging, mainly because : (i) data items are rich and heterogeneous: experiment details, raw data, scientific interpretations, images, literature, etc. ii) data items are distributed over many heterogeneous data sources rendering a complex integration, iii) data are speculative and subject to errors and omissions within these data sources, and bio-data quality is difficult to evaluate, and iv) bio-medical knowledge is constantly morphing and in progress..

This paper reports on our experience in building GEDAW: an object-oriented Gene Expression Data Warehouse to store and manage relevant information for analyzing gene expression measurements [12]. GEDAW (Gene Expression DAta Warehouse) aims on studying in silico liver pathologies by using expression levels of genes in different physiopathological situations enriched with annotations extracted from the variety of the scientific sources and standards in life science and medicine.

A comprehensive interpretation of a single gene expression measurement requires the consideration of the available knowledge about this gene, including its sequence and promoters, tissue-specific expression, chromosomal location, molecular function(s) and classification, biological processes, mechanisms of its regulation, expression in other pathological situations or other species, clinical follow-ups and, increasingly important, bibliographic information. Beyond the process of data clustering, this knowledge provides representations that can help the scientist to address more complex questions and suggest new hypothesis, leading in our context to a clearer identification of the molecular regulation mechanisms involved in the occurrence of liver diseases and at mid-term to new diagnostic and therapeutic tools.

The required knowledge is spread world-wide and hosted on multiple heterogeneous resources. Manually navigating them to extract relevant information on a gene is highly time-comsuming and error-prone. Therefore, we have physically integrated into GEDAW a number of important sources in life science and medicine that are structured or semi-structured. Our final objective is to propose a more systematic approach to integrate data on liver genes and to organize and analyze them within a target question - which is in our case specific to an organ and a pathological state. This is a complex task, with the most challenging questions being: i) bio-knowledge representation and modeling, ii) semantic integration issues and iii) integrated bio-data analysis.

Building a scientific data warehouse to store microarray expression data is a well studied problem. Conceptual models for gene expression are for instance discussed in [18].The Genomic Unified Schema (GUS) integrates diverse life science data types including microarray data, and a support of data cleansing, data mining and complex queries analyses, thus making it quite generic [2]. The warehouse of [11] focuses on storing as possible details on the experiments and the technologies used. In GEDAW we only focus on the result of an experiment, i.e., expression measurements. No further experimental details are stored within the warehouse. The Genome Information Management System (GIMS) in which one of the authors has been participating, allows the storage and management of microarray data on the scale of a genome, making GIMS, in contrast to GEDAW, a genome-centric rather than gene-centric data warehouse [9]. Finally, [10] describe the GeneMapper Warehouse for expression data

integrating a number of genomic data sources. In contrast, GEDAW has a focus on medical and "knowledge-rich" data sources.

1.1 Architecture for BioData Integration

GEDAW is a gene-centric data warehouse devoted to the study of liver pathologies using a transcriptome approach. New results from medical science on the gene being studied are extremely important to correlate gene expression patterns to liver phenotypes. To connect to this information, we take advantage of the recent standards developed in the medical informatics domain, i.e., the UMLS knowledge base. [3]

GEDAW schema includes three major divisions: (i) gene and gene features along with transcripts and gene products division, (ii) expression measurements of liver genes division generated by in-house experiments and (iii), universal vocabularies and ontologies division. As illustrated in Figure 1, to store the gene expression division a local relational database has been built, as a repository of array data storing as many details as possible on the methods used, the protocols and the results obtained. It is a MIAME (Minimum Information About Microarray Experiment) compliant source [6].

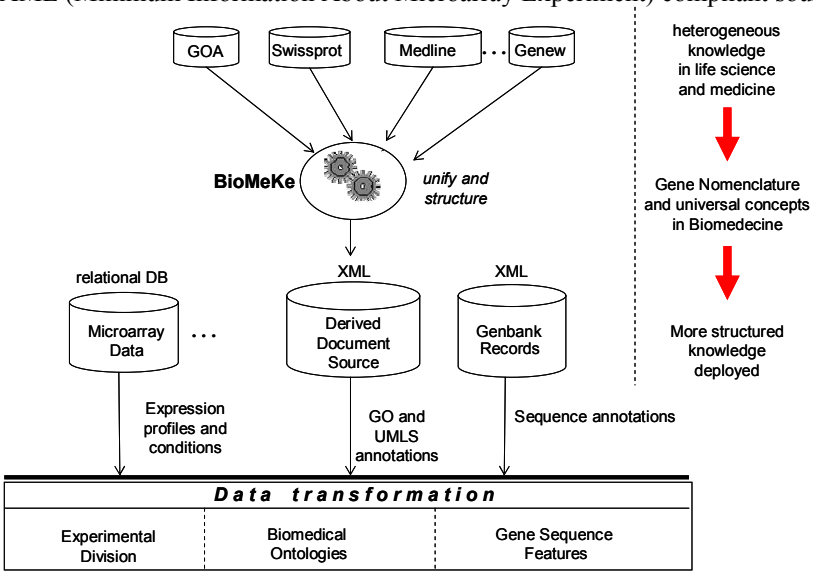

Fig. 1. GEDAW System Architecture

The sources currently integrated are spread world wide and hosted on different representation systems, each having its own schema. XML records from the GENBANK [7] have been used to populate the gene sequence features division into GEDAW.

Explicit relationships associating genes and their expression profiles with diseases are also extremely needed to understand the pathogenesis of the liver. For this purpose, we use the system BioMeKE [8,17] to curate the ontology division of each expressed gene with relative concepts in life science and medicine. The BioMEdical Knowledge Extraction module (BioMeKE) includes the Unified Medical Language

System® (UMLS) covering the whole biomedical domain, and the Gene Ontology™ (GO) that focuses on genomics. It includes additional terminologies, as that provided by the HUman Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) to resolve synonymy conflicts [19]. An XML document that annotates each gene by exploring these biomedical terminologies is derived from BioMeKE. It is then parsed and integrated into the warehouse.

1.2 Contribution

The aim of this paper is to share our experience on designing and implementing an integration process for biomedical data in the presence of syntactic and semantic conflicts. Other aspects such as biological data quality controlling, mining and refreshing will be described elsewhere.

1.3 Outline

An overview on the biological background and the questions that motivate the design of GEDAW are given in the next section. In section 3, the provenance, content and the format of the structured resources used for integration in GEDAW are described. In section 4, the integration process along with a brief schema design is presented. The data mapping rules that have been defined for instances conciliation and cleansing during the integration process are also presented. The generic interface used for queries composition and execution is tackled in section 5. Section 6 concludes and presents the perspectives of our future works.

2 Biological Background and Motivations

Transcriptome is the study of the transcriptional response of the cell to different environment conditions such as, growth factors, chemicals, foods treatments, genetic disturbance, etc. The cell may response by an excessive expression or repression of certain genes in two different situations, for example normal vs. pathologic.

2.1 Transcriptome experiments

In the liver framework, the objective of transcriptome experiments is to emphasize both co-expressed genes and gene networks in a specific pathology within the hepatocyte.

To determine whether a single gene is expressed is a routine task for a biologist, but this process becomes more complicated because the data generated are massive. DNA-chips are indeed used and thousands of genes are deposited on a two dimensional grid. The experiment generating thousands of data points requires an efficient processing of the storage and the management of data. The key question is: which of (and why?) the deposited genes are abnormally expressed in the injured tissues? Each

gene is represented by a spot, and its expression level is measured by means of the spot intensity. This same gene does have other multiple features, recorded in World Wide Web resources, and that must be considered to answer such questions.

2.2 Biomedical Issues Underlying Data Integration

To study experimental data, the scientist expects an integrated environment that captures his own experimental data enriched with information and expertise on the expressed genes. Beyond the process of clustering expression measurements in gene clusters, such an integrated environment should allow him to better focus on the scientific interpretation derived from such a clustering that reveals such clusters.

Together with the collected gene data, the integrated environment should be able to answer questions that need an integration of knowledge from the biological level to the pathological level. Below we give three types of questions that scientists frequently ask and that cannot be answered by simple SQL queries, but require the application of data mining techniques.

1 The set of genes that have seen their expression modified in a given condition?
2 Within this set, is there a subset of genes that are co-regulated?
3 What are the elements that may explain a parallel (or opposite) modulation of certain genes: membership to a functional class, homologies occurring in their peptides sequences, or in their nucleic sequences particularly in the promoting region?

Scientists may need to go thoroughly into sequences (question 3.) of the co-expressed genes for discovering common motifs, because genes sharing similar expression profiles must share transcription regulation mechanisms that include common transcription factors. They also need to go thoroughly into disease information and clinical follows-up in order to find out correlations between particular mutants' phenotypes and expression patterns. The integrated environment should also be able to answer questions such as:

1 Is there any correlation between gene expression levels and a certain pathological phenotype?
2 What is the set of genes for which a dysregulation characterizes a pathological sample by indicating a gravity level, a prognostic factor, a sensitivity level or on a contrary a resistance to a certain treatment ?

Respective genes annotations that comes from the UMLS knowledge-base and the Gene Ontology, along with gene expression profiles, are used to proceed such questions. Relative conceptual terms in both ontologies are extracted from the unified document-source, derived by BioMeKE.

2.3 GEDAW: An Object-Oriented Environment for Integrating Liver Genes Data

Considering the different integration issues previously described, an object oriented data warehouse called GEDAW (Gene Expression DAta Warehouse) has been designed for integrating and managing : i) data being produced on the expressed genes

in public databanks and literature, ii) normalized experimental data produced by Microarray experiments and iii) complementary biological, genomic, and medical data.

3 Data Resources

Searching across heterogeneous distributed biological resources is increasingly difficult and time-consuming for biomedical researchers. Bioinformatics is coming to the forefront to address the problem of drawing effectively and efficiently information from a growing collection of multiple and distributed databanks. Several resources can be used to instantiate the liver warehouse GEDAW. We describe here the ones that have been selected for having the most appropriate properties, enabling a systematic extraction of gene attributes: 1) experiment resources, 2) genomic databanks and 3) ontological resources. We demonstrate for each selected resource, its provenance, content, structure and which gene attributes are being extracted.

3.1 Experimental Resources

To not burden the warehouse, a MIAME compliant relational database has been built independently (Figure2), in order to store and manage experimental microarray data [12]. This database stores as much as possible details on the microarray experiments, including the techniques used, protocols, samples and results obtained (ratios and images).

We will not go in further details concerning this database, except saying that it acts as a permanent source of expression levels delivered by in-house transcriptome experiments on injured liver tissues, and provides facilities to select and export data. Part of those data is exported to the data warehouse.

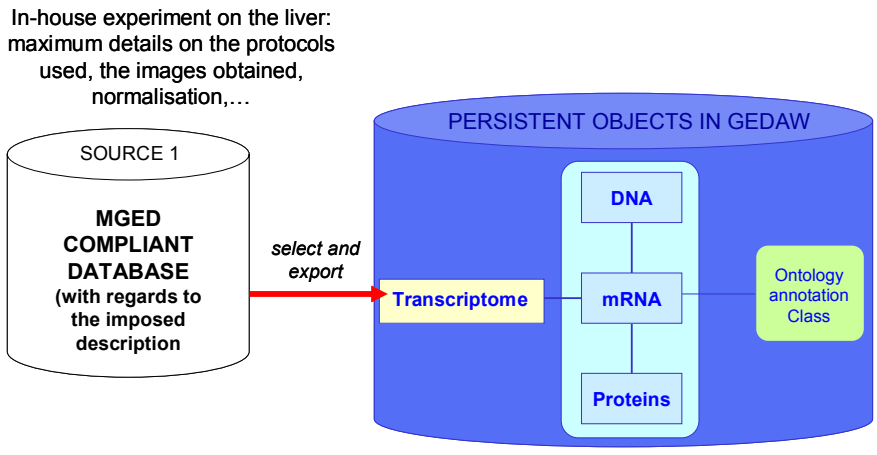

Fig. 2. An external source to manage liver transcriptome experiments

3.2 Genomic Databanks Resources

In order to perform consistent analyses on the expressed genes, the integration of the precise pre-existing annotations of their sequences is necessary. Sequence data to consider include: 1) the DNA sequence and sequence components : known promoters, known transcription binding sites, introns, exons, known regulators, 2) the mRNA sequence, sequence components and alternative transcripts and 3) functional proteins. Being conscious that an exhaustive gene annotation is available for a limited number of genes, it is however helpful to infer new knowledge on yet unknown co-expressed genes.

Data describing genomic sequences are available in several public databanks via Internet: banks for nucleic acids (DNA, RNA), banks for protein (polypeptides, proteins) such as SWISS-PROT , generalist or specialized databanks such as GENBANK , EMBL (European Molecular Biology Laboratory), and DDBJ (DNA DataBank of Japan). Each databank record describes a sequence with its several annotations.

As an example, the description of the Homosapiens Hemochromatosis gene HFE, which mutation causes a genetic liver disease having the same name is given in GENBANK. The description of this gene is available in both HTML[2] and XML[3] formats. An XML format that focused on the sequence of HFE gene is also available[4].

Each record is also identified by a unique accession number and may be retrieved by key-words. Annotations include the description of the sequence: its function, its size, the species for which it has been determined, the related scientific publications (authors and references) and the description of the regions constituting the sequence (start codon, stop codon, introns, exons, ORF, etc.). GENBANK (with more than 20 million records of different sequences) [7] is one of the first banks that propose XML format for its records with a well-defined DTD specifying the structure and the domain terminology for the records of genes and submitted sequences.

3.3 Ontological Resources

Relating genotype data on genes with their phenotype during the integration process is essential to be able to associate gene expression levels to a pathological phenotype.

Tremendous web resources provide such information for a given gene. But their heterogeneity is a major obstacle for a consistent semantic integration. They are numerous and continually evolving, the number of biomolecular entities is very large, the names of biological entities are associated with synonymy: a gene can have multiple aliases (synonyms) in addition to its official symbol, and genes that are functionally different across species may have the same name (ambiguity) [14,20], different databases organize data according to different schemas and use different vocabularies. Shared ontologies are used to conciliate and to attain as much as possible data conflicts. Various standards in life science have been developed to provide domain knowledge to be used for semantically driven integration of information from different sources.

---

[2] www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=1890179

[3] www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&list_uids=1890179&dopt=xml

[4] www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&list_uids=1890179&dopt=gbx

3.3.1 Gene Ontology

Gene Ontology™ (GO) is an ontology for molecular biology and genomics [13]. The three hierarchies of GO are molecular function (F), biological process (P) and cellular component (C). GO terms are used as attributes of gene products to provide information about the molecular functions, the biological processes, and the cellular components related to the gene product. In our context of high throughput transcriptome experiments, we use GO to annotate the genes expressed in different situations in the liver. Furthermore, GO is broadly used by public databanks to annotate genes. Therefore, it has become a standard and plays an important role in biomedical research, by making possible to draw together information from multiple resources. To illustrate with an example, to the ceruloplasmin concept (a gene involved in iron transport, having a central role in iron metabolism and is secreted in plasma by hepatocytes) is associated the set of concepts in each hierarchy of GO ontology (Table 1).

Molecular function	Biological process	Cellular Component
Multicopper Feoxidase iron Transport mediator	Iron homeostasis	Extracellular space

Table 1. Ceruloplasmin annotations in Gene Ontology

3.3.2 UMLS Knowledge Base

The UMLS is developed by the US National Library of Medicine. It comprises two major inter-related components: the Metathesaurus®, a large repository of concepts (around 900,000 concepts), and the Semantic Network, a limited network of 135 Semantic Types [3]. The Metathesaurus is built by merging existing vocabularies, including Medical Subject Headings (MeSH), which is used to index biomedical literature in MEDLINE, and GO. In the Metathesaurus, synonymous terms are clustered under a same concept, each having a Concept Unique Identifier (CUI). To the ceruloplasmin concept is associated the CUI:C0007841 and a set of synonymous terms (Table 2a) (2003AC release of the UMLS).

Although the UMLS was not specifically developed for bioinformaticists, it includes also terminologies such as the NCBI taxonomy, OMIM terminology and GO that are of great interest for biologists. It also includes the MeSH, which is used to index MEDLINE abstracts. Therefore, the UMLS is a means to integrate resources since it integrates (repetition) terminologies that are used to represent data in various resources. The second motivation is that the UMLS contains 12 million relations among the Metathesaurus concepts. The source vocabularies provide hierarchical relations. RO (Other Relation) relations associate concepts from different kinds, such as diseases and tissues, or diseases and kinds of cells. In addition, co-occurrences in MEDLINE are also represented in the UMLS [3]. The last motivation is that the UMLS includes an upper level ontology of the biomedical domain (the UMLS Semantic Network) made of 135 Semantic Types. Each Metathesaurus concept is assigned to one or more Semantic Types. Three major relations are then concerned and extracted for each concept from UMLS :

- Parent concept (Table 2b): the parents of ceruloplasmin concept illustrate hierarchical relations in UMLS.
- Related concepts in diseases (Table 2c), tissues or kind of cells.

- Co-occurrences in Medline concepts (Table 2d), each with an additional numeric frequency.

Synonymous	Parents concepts	Related concepts	Co-occurred Concepts in MEDLINE
Ceruloplasmin alpha(2)-Ceruloplasmin Ceruloplasmin Ferroxidase Ceruloplasmin Oxidase CP - Ceruloplasmin Fe(II):oxygen oxidoreductase ferroxidase <1>	Alpha-Globulins Acute-Phase Proteins Carrier Proteins Alpha-Globulins Metalloproteins Oxidoreductases Enzyme	Copper Menkes Kinky Hair Syndrome copper oxidase Serum Ceruloplasmin Test Ceruloplasmin Serum Decreased Ceruloplasmin measurement	Copper Iron Antioxidants Hepatolenticular Degeneration Ferritin Brain Liver Superoxide Dismutase
(a)	(b)	(c)	(d)

Table 2. Ceruloplasmin annotations extracted from UMLS

3.3.3 Other Resources: Terminologies

At present, an additional terminology is mainly used to manage heterogeneity in naming genes, gene products or diseases, as well as in identifying items in different databanks. Given a term or a gene symbol, lexical knowledge is needed to deal with synonyms and find the corresponding concept. Available resources in the biomedical domain include the Genew database developed by the Human Gene Nomenclature Committee to provide approved names and symbols for genes, as well as previous gene names and symbols [19].

3.3.4 Mapping Ontologies into GEDAW

The use of ontologies and terminologies terms as attributes values for genes has been made possible by the joint application project BioMeKE [17]. A local consistent support into BioMeKE system of the terminologies described above enables the extraction of respective nomenclature and conceptual terms in biology and medicine, given a gene name, a symbol, or any gene relative identifier in biomedical databanks. To navigate through these resources, a set of JAVA functions have been developed to:

- Find all the synonyms of a term and all the identifiers of a gene or gene product in Genew and the UMLS Metathesaurus,
- Provide the cross-references between a gene and a protein (e.g. SWISS-PROT ID) from Genew.
- Represent the different paths to reach the information about a gene or a gene product via all the available cross-references.
- Search for information about a gene or a gene product, i.e. the set of concepts related to this gene in GO (molecular function, biological process and cellular component) and the set of concepts related to the gene in UMLS including chemicals and drugs, anatomy, and disorders.

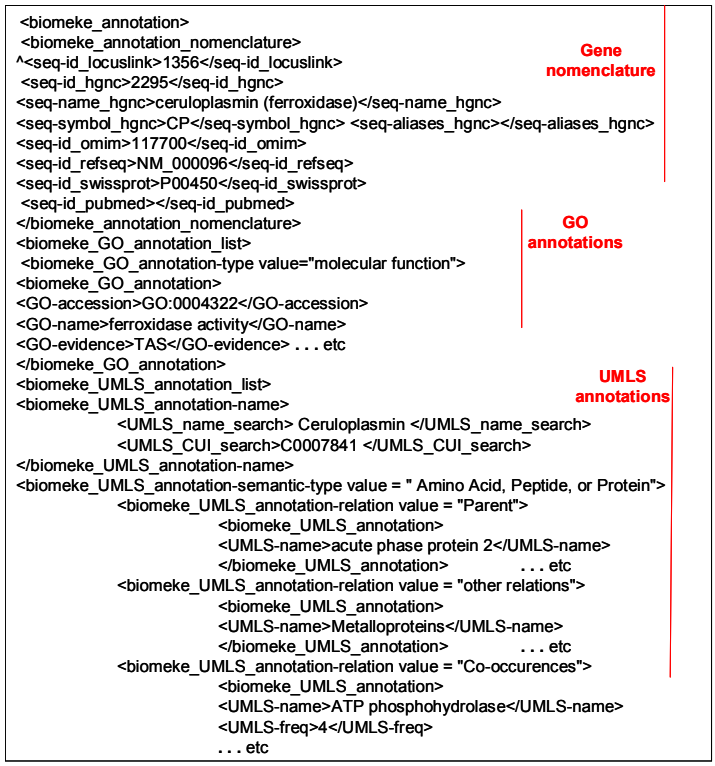

Fig. 3. BioMeKE-xml document to annotate the ceruloplasmin Gene

These annotations are then considered by the expert, filtered and stored within the warehouse for further classifications using gene expression profiles. Because the aim of this paper is not to describe BioMeKE but rather to introduce its general scope and outputs, we will not go in further details. We suggest the reader to get further details in another paper devoted to this application [8,17].

To annotate each expressed gene, BioMeKE delivers an XML document (Figure 3) to be parsed, transformed and stored into GEDAW within the Ontology_annotation Class. This document-source standing as a structured data source derived by BioMeKE.

4 Bio-Data Integration

Designing a single schema that integrates syntactically and semantically the whole heterogeneous life science data sources is still a challenging question. Integrating the source schemas is presently the most commonly used approach in literature [15]. By restricting ourselves to structured or semi-structured data sources, we have been able to use a schema mapping approach with the GAV paradigm [16]. In our context, schema mapping is the process of transforming data conforming to a source schema to the corresponding warehouse schema by the definition of a set of mapping rules. The

data sources include : i) GENBANK for the genomic features of the genes recorded in XML format, ii) conceptual annotations derived from the biomedical ontologies and terminologies using BioMeKE outputs as XML documents, iii) and gene expression measurements selected from the in-house relational database.

By using a mapping approach from one source at a time, we have minimized as much as possible the problem of identification of equivalent attributes between sources, whereas the problem of duplicate detection is still important. Identifying identical objects in the biomedical domain is a complex problem, since in general the meaning of "identity" cannot be defined properly. In most applications, even the identical sequences of two genes in different organisms are not treated as a single object. In GENBANK, each sequence is treated as an entity in its own, since it was derived using a particular technique, has particular annotation, and could have individual errors. For example, there are more than 10 records for the same DNA segment of the HFE gene. Thus, classical duplicate detection methods [22] do not suffice. Duplicate detection and removal is usually performed either using a simple similarity threshold approach, as in the case of GEDAW, or based on manual intervention for each single object, such as in RefSeq. Data submission to public biological databanks is often a rather unformalized process that usually does not include name standardization or data quality controls. Erroneous data may be easily entered and cross-referenced. Even if a tool like LocusLink[5]proposes a cluster of records, across different biological databanks, as being semantically related, biologists still must validate the correctness of the clustering and resolve value differences among the records.

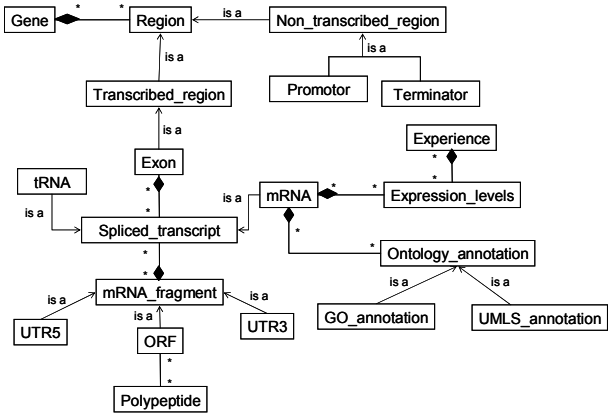

Fig. 4. GEDAW UML Conceptual schema

In GEDAW, a unique schema (Figure 4) has been defined to describe different aspects of a gene, to which has been added an ontological annotation class associated to each gene transcript. The stored ontological annotations represent the more specialized concepts associated to the genes. The ontology annotation class used for storing the terms from both medical and biological terminologies includes attributes like: ontology and annotation type along with category, value and description attributes of a term. These attributes are extracted by parsing the XML files delivered by BioMeKE.

At the schema-level, the problem of format heterogeneity makes necessary to

[5] www.ncbi.nlm.nih.gov/LocusLink

transform data, so that they conform to the data model used by our warehousing system. Information sources consist of sets of XML files, while the GEDAW target schema is object-oriented. This translation problem is inherent in almost all data integration approaches, but becomes much more complex in the biological domain because the potentially different (and not formalized yet) biological interpretations of schema elements and the fact that, together with the current state of knowledge, schemas and interpretations tend to evolve quickly and independently in the different sources.

In order to define an appropriate data aggregation of all the available information items, data conflicts have to be resolved using rules for mapping the source records and conciliating different values recorded for a same concept. Mapping rules are defined to allow the data exchange from the public databanks into GEDAW (Figure 5). Apart from experimental data, public information items are automatically extracted by scripts using the DTD (Document Type Definition) of the data source translated into the GEDAW conceptual data model.

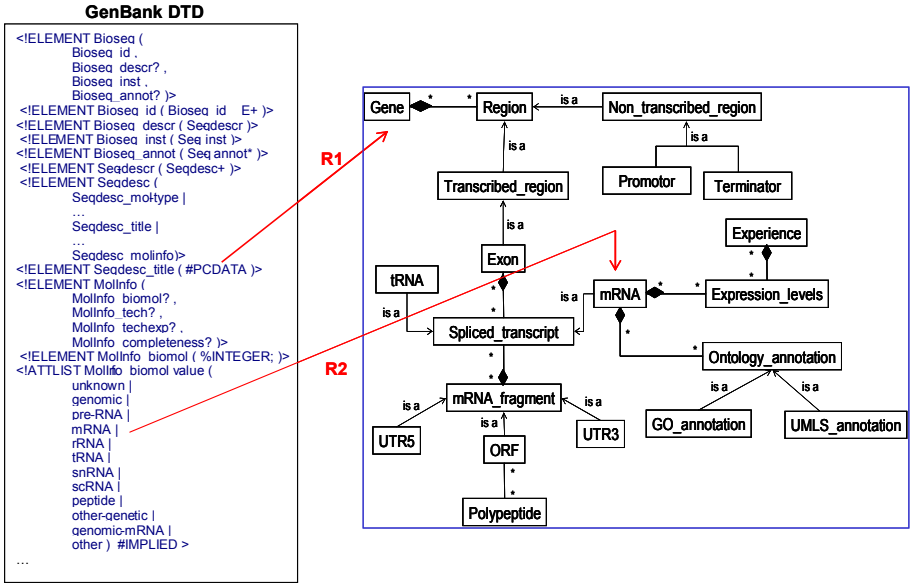

Fig. 5. Example of mapping rules between GENBANK DTD and GEDAW schema

Three categories of mapping rules are proposed: 1) structural mapping rules, 2) semantic mapping rules and 3) cognitive mapping rules according to the different knowledge levels and perspectives for biological interpretation.

The structural mapping rules are defined at the schema level according to the GEDAW model by identifying the existing correspondences with relevant DTD elements (e.g., the *Seqdesc_title* element in GENBANK DTD is used to extract the name "name" of the gene and the *MolInfo_biomol* value its type of molecule with respectively structural mapping rules R1 and R2 in Figure 5). Then, the records of interest are selectively structured and data are extracted.

Semantic and cognitive mapping rules are used for data unification at the instance level: several rules may use available tools for determining analogies between ho-

mologous data (such as sequence alignment, for example): the result of the BLAST algorithm (implemented in a set of similarity search programs for Basic Local Alignment Search Tool) allows considering that two sequences match. The nomenclature section provided by BioMeKE (Figure 3) is also considerably used to conciliate duplicate records. More semantic mapping rules have been built using this information during the process of integration. For example, the Locus-ID is used to cluster submitted sequences associated to a same gene (cross-referenced in LocusLink) and the official gene name along with its aliases to relate different gene appearance with different names, in literature for example.

Let us consider three distinct selectively structured records we may obtain from GENBANK databank by querying the DNA sequence for gene HFE. A first record identified by the accession number AF204869 describes a partial sequence (size = 3043) of the HFE gene with no annotation but one relevant information item about the position of the promoter region. A second record identified by the accession number AF184234 describes a partial sequence (size = 772) of the protein precursor of HFE gene with a detailed but incomplete annotation. The third record identified by the accession number Z92910 describes the complete sequence (size = 12146) of the HFE gene with a complete annotation. In this example, BLAST(sequence(Z92910), sequence(AF184234))=100% indicates the sequence in both records are perfectly homologous and can be merged. Cognitive mapping rules may be used in this example for conciliating data such as:

R3 : Descriptive Inclusion: record(Z92910) contains record(AF184234)
R4 : Position Offset: position(Z92910.exon)=6364+position(AF184234.exon)

In our context a liver cDNA microarray corresponding to 2479 cDNA clones spotted onto glass slides has been designed. The data unification process described above has lead to identify 612 distinct genes on the 2479 deposited clones. A complete integration of 10 hybridization experiments took around one day runtime, with around 11 Mbytes charged database size.

5 Integration Results Construction and User Interface

Now to recapitulate, the integration process of transcriptomic data into GEDAW is operated in four steps. During the first step, to the probes (or clones) used by in-house experiments, is associated a set of gene names, in terms of accession numbers of similar sequences in GENBANK along with textual descriptions. The second step is in charge of selecting the set of experiments for which the researcher wishes to integrate and analyse the experiments results, and then of loading expression levels measured for these genes. For each gene having its expression levels in different physiopathological situations already stored in GEDAW, the full annotation of the sequence associated to this gene is loaded from GENBANK by XML transformation to Objects. BioMeKE is launched in Step 4 to bring for each integrated gene its nomenclature and its ontological annotations in life science from Gene Ontology and in medicine from UMLS. In step 5, the results are delivered to the expert, for a filtering phase using either predefined mapping rules, output nomenclature, or simply his expertise, to eliminate duplicate records of genes.

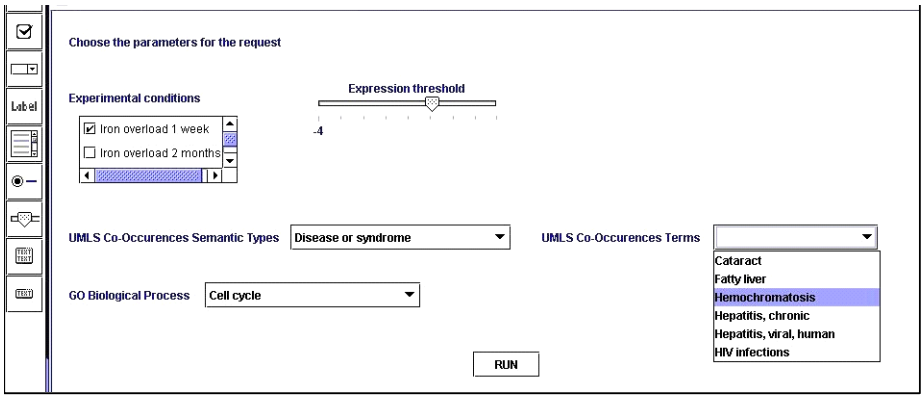

Fig. 6. Example of Query Composition

When the user poses a query, the whole integration results for each gene are brought in. Further refinements on these data can be operated, by selecting for example genes having expression levels between a minimum value and a maximum value, those belonging to a given biological process or co-occurring in Medline with a given concept, or having a known motif in their mRNA sequences and co-located on a same chromosome. It could be also a conjunction of these criteria. In Figure 6, we show an example of a query composed in the generic java-based interface we have developed for GEDAW. Resulting sets are presently browsed using either FastObjects interface, or delivered as Textfiles to the expert for further analyses.

6 Conclusion

The GEDAW system presented in this paper allows massive importation of biological and medical data into an object-oriented data warehouse that supports transcriptome analyses specific to the human liver. This paper focused on the relevant genomic, biological and medical resources that have been used to build GEDAW. The integration process of the full sequence annotations of the genes expressed is described. It is performed by parsing and cleaning the corresponding XML description in GENBANK, transforming the recorded genomic items to persistent objects and storing them in the warehouse. This process is almost systematic because another aspect related to the conciliation of duplicate records has been added. Elements of formalization of expertise rules for mapping such data were given. This ongoing work is still a difficult problem in information integration in life science and has not yet satisfied answers by classical solutions proposed in existing mediation systems.

In order to lead strong analysis on expressed genes and correlate expression profiles to liver biology and pathological phenotype, a second way of annotation has been added to the integration process. We chose to integrate Gene Ontology, due to its available biological annotations in the most used bio-computer resources, mainly Swissprot, GENBANK, Ensembl, TrEMBL and LocusLink databanks. It is also referenced in other relevant ontologies, like MGED [21]. More important is our considera-

tion during integration of the medical annotations of the genes from UMLS, a well considered knowledge base in Medical Informatics [3,4,5]. These ontological annotations have been delivered by BioMeKE within the semi-structured document source BioMeKE-xml. Also, because a gene may have different appearances with different names in several bio-data banks and literature the approved nomenclature of the gene and its synonyms have been collected in BioMeKE-xml. This information is also a pre-requisite to resolve the problem of duplicate records.

An exhaustive integrated tool that facilitates access to diverse data on the expressed genes is then provided to the researcher. Intensive querying of the integrated database using OQL queries has been conducted with multiple criteria on genes attributes. Current investigations are focusing on the application of advanced data mining techniques for a combined analysis of expression levels on genes with enriched annotations, and functional similarities are likely to reveal authentic clusters of genes.

With regards to the limits of our warehousing approach, it is relevant as long as data integration from the heterogeneous sources in Biomedicine and their refreshment in the warehouse stay feasible automatically and with a reasonable performance. One argument in favor of actually storing data in GEDAW instead of dynamically linking to the corresponding sources concerns reproducibility purposes, i.e., being able to analyze several gene expression data in reference to the same domain knowledge at different times. BioMeKE system provides domain knowledge useful for acquiring information from diverse resources. It is intended to be an ontology-based mediation system that continuously supplies the gene expression warehouse with a homogeneous access to multiple data sources in Biomedicine. A filtering task is nevertheless performed by the expert on the delivered annotations before their storage in the warehouse by using multiple criteria, like the frequency information of a concept co-occurrences in Medline.

The standard ontologies such as GO and UMLS continue to evolve. They are physically supported by BioMeKE system rather than accessed via the web, making possible their refinement to expert knowledge in specific sub-domains like the liver or the iron metabolism. An interesting point to quote is the acquisition of news concepts and relationships from the analyses operated on the transcriptome data. Expressive and formal representation of this new biomedical knowledge will then be gradually added to the domain, allowing the expansion of queries on transcriptomic data.

Acknowledgements: This work was supported by grants from Region Bretagne (20046805) and inter-EPST. Emilie Guérin was supported by a MRT fellowship and grants from Region Bretagne.

References

[1] Achard, F., Vaysseix, G. and Barillot, E. (2001) XML, bioinformatics and data integration, Bioinformatics, 17(2), 115-125.

[2] Babenko V, Brunk B, Crabtree J, Diskin S, Fischer S, Grant G, Kondrahkin Y, Li L, Liu J, Mazzarelli J, Pinney D, Pizarro A, Manduchi E, McWeeney S, Schug J, Stoeckert C.(2003) GUS The Genomics Unified Schema A Platform for Genomics Databases. http://www.gusdb.org/

[3] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
[4] Bodenreider O, Burgun A. Aligning Knowledge Sources in the UMLS: Methods, Quantitative Results, and Applications. Medinfo. 2004;2004:327-31.
[5] Bodenreider O, Mitchell JA, McCray AT. (2002) Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. Proc AMIA Symp. 2002; : 61-5.
[6] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.
Nat Genet. 2001 Dec;29(4):365-71.
[7] Benson D.A, Karsch-Mizrachi I, Lipman D.J, Ostell J, and Wheeler D.L. GENBANK: update, Nucleic Acids Res., Jan 2004; 32: 23 - 26.
[8] Burgun A, Bodenreider O, Le Duff F, Moussouni F, Loréal O. Representation of roles in biomedical ontologies : a case study in functional genomics. JAMIA (supl), Proc. AMIA 2002 Symp, 86-90
[9] Cornell M, Paton NW, Wu S, Goble CA, Miller CJ, Kirby P, Eilbeck K, Brass A, Hayes A, Oliver SG (2001) GIMS - a data warehouse for storage and analysis of genome sequence and functional data. Proc. 2nd IEEE International Symposium on Bioinformatics and Bioengineering (BIBE) 15-22.
[10] Do, H.-H. and Rahm, E. (2004). "Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach". EDBT'04, Heraklion, Greece, Springer LNCS.
[11] Fellenberg K, Hauser N.C, Brors B, Hoheisel J.D, and Vingron M. Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis, Bioinformatics, Mar 2002; 18: 423 - 433.
[12] Guerin E., Marquet G., Moussouni F., Burgun A., Mougin F., Loréal O. Deployment of heterogeneous ressources of genomic, biological and medical knowledge on the liver to build a datawarehouse. Proc. ECCB 2003, pp. 59-60
[13] Harris MA et. al. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D258-61.
[14] Kashyap V, Sheth A. (1996) Schematic and semantic similarities between database objects: a context –based approach. Int. J. Very Large Data Bases, 5(4): 276-304
[15] Lakshmanan L, Sadri F, Subramanian I, : On the logical Foundation of Schema Integration and Evolution in Heterogeneous Database Systems. DOOD International Conference (1993) 81-100
[16] Maurizio Lenzerini. Data integration: a theoretical perspective. In Proc. of PODS 2002.
[17] Marquet G, Burgun A, Moussouni F, Guerin E, Le Duff F, Loreal O. BioMeKE: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis. Stud Health Technol Inform. 2003;95:80-5.
[18] Paton N.W, Khan S.A, Hayes A, Moussouni F, Brass A, Eilbeck K, Goble C.A, Hubbard S.J, and Oliver S.G. Conceptual modelling of genomic information, Bioinformatics, Jun 2000; 16: 548 - 557.
[19] Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. (2001) The HUGO Gene Nomenclature Committee (HGNC).Hum Genet.;109(6):678-80
[20] Tuason O, Chen L, Liu H, Blake JA, Friedman C.(2004) Biological nomenclatures: a source of lexical knowledge and ambiguity. Pac Symp Biocomput. 2004;:238-49.
[21] MGED Microarray Gene Expression Data (MGED). A guide to microarray experiments--an open letter to the scientific journals. Lancet. 2002 Sep 28;360(9338):1019
[22] Galhardas, H., Florescu, D., Sasha, D., Simon, E. and Saita, C.-A. (2001). "Declarative Data Cleaning: Model, Language, and Algorithms". 27th Conference on Very Large Database Systems, Rome, Italy.