Combining biomedical knowledge and transcriptomic data to extract new knowledge on genes

Emilie Guérin[1,§] and Gwenaëlle Marquet[2,§], Julie Chabalier[2], Marie-Bérengère Troadec[1], Christiane Guguen-Guillouzo[1], Olivier Loréal[1], Anita Burgun[2] and Fouzia Moussouni[1]

[1]INSERM U522, IFR 140, Université de Rennes 1, CHU Pontchaillou, 35033 RENNES Cedex, France

[2]EA 3888, IFR 140, Université de Rennes 1, Faculté de Médecine, 35043 RENNES Cedex, France

Abstract

In biomedical research, interpretation of microarray data requires confrontation of data and knowledge from heterogeneous resources, either in the biomedical domain or in genomics, as well as restitution and analysis methods adapted to huge amounts of data. We present a combined approach that relies on two components: BioMeKE annotates sets of genes using biomedical GO and UMLS concepts, and GEDAW, a Gene Expression Data Warehouse, uses BioMeKE to enrich experimental results with biomedical concepts, thus performing complex analyses of expression measurements through analysis workflows. The strength of our approach has been demonstrated within the framework of analysis of data resulting from the liver transcriptome study. It allowed new genes potentially associated with liver diseases to be highlighted.

1 Introduction

New high throughput technologies used to study transcriptome, proteome or metabolome, produce large amounts of data. The exploitation of these data requires important database solutions to manage experiment results with relevant information, including functional annotations and gene-disorder relations, as well as data mining techniques to extract new knowledge[1, 2].

In the case of transcriptome study, a comprehensive interpretation of a single gene expression measurement requires the consideration of all available knowledge on this gene, including: i) its genomic annotations, *i.e.* he chromosomal localization of the gene and related sequences, ii) the biological knowledge, *i.e.* he biological processes in which the gene is involved and the target functions in these processes, and iii) the medical knowledge, *i.e.* the different symptoms, syndromes and diseases associated to the gene. A comprehensive representation of this knowledge can help scientists to address more complex questions and suggest new hypotheses, leading to a clearer identification of the molecular and biological mechanisms involved in the diseases.

Interrelating the different kinds of information about genes is challenging as data are spread over the web, hosted in a large scale of independent, heterogeneous and highly focused resources [3, 4]. Moreover, within those sources, biological data are complex, often redundant and complementary. In this context, creating an infrastructure for a unified biological knowledge resource is a key to an effective and accurate analysis of transcriptomic data.

---

§ Corresponding authors, who contributed equally to this work

Integration of biological databases has been an ongoing research problem and several approaches have been proposed [4]. Among those approaches, data warehouse systems, by materializing the data from multiple sources into a local environment, allow to improve the efficiency of query optimization, as such corresponding to the most adapted systems to analyse large results sets as obtained by microarrays [5]. Several data warehouses devoted to transcriptome analysis have been developed. Examples are GIMS [6], M-Chips [7], GenMapper [8] and GeWare [9]. However, most existing methods do not combine medical knowledge with genomic and biological information.

Standard biomedical vocabularies have been developed in both domains (biological and medical). Gene Ontology (GO[1]) is a controlled vocabulary for molecular biology and genomics [10] useful to annotate gene products in most public databanks. However, GO does not provide information on pathologic conditions and disorders that have been associated with genes and their products. The Unified Medical language System® (UMLS[2]) covers the whole biomedical domain and is intended to help health professionals and researchers by merging more than 100 vocabularies [11]. While GO is widely used to provide functional annotations of gene products the UMLS is mainly used in medical informatics. The UMLS appears to be a potential resource for providing associations between genes and medical knowledge, which may complement GO annotation.

With regards to transcriptome, raw expression data are not sufficient to carry out an exhaustive analysis, since the result would be clusters of genes sharing expression profiles that need to be interpreted. To go beyond clusters, our challenge was to combine experimental data both with genomic data and biomedical knowledge and then to mine transcriptomic data under an expert supervision.

We have combined: i) BioMeKE[3] (Biological and Medical Knowledge Extractor), a system that supports the GO and the UMLS vocabularies to annotate any sets of genes with biomedical concepts and ii) GEDAW, a Gene Expression DAta Warehouse that integrates microarray experimental results enriched with multiple complementary information extracted from web sources, thus performing complex analyses of expression measurements through analysis workflows [12]. BioMeKE provides biomedical annotations based on GO and the UMLS for a set of genes and GEDAW integrates these biomedical annotations with genomic data. In this paper, we demonstrate that we were able to learn about the genes by mining combined data and concepts from both systems, in a context of biomedical research and expert supervision.

After a presentation of BioMeKE and the data mining method used for enriching transcriptomic data in section 2, we describe, in section 3, its usage for generating biomedical knowledge on genes expressed in different physiopathological situations in the liver, within an expert guided data mining approach. We conclude in section 4 with a discussion demonstrating strong interest of our work in the context of liver diseases study and also its limits.

[1] http://www.geneontology.org
[2] http://www.nlm.nih.gov/research/umls/
[3] http://www.med.univ-rennes1.fr/biomeke/

2 Methods

2.1 Biomedical annotation

BioMeKE (Biological and Medical Knowledge Extraction system) is a system based on two standard terminologies, Gene Ontology (GO) that focuses on molecular biology and genomics, and the Unified Medical Language System (UMLS) that covers the whole biomedical domain. To deal with heterogeneity in naming genes, the system includes also the main gene nomenclature database, Genew.

GO is the main biological controlled vocabulary widely used to describe the genes so to have a view of their main functions. GO contains 18,735 terms (May 2005) organized through a Direct Acyclic Graph (DAG) divided in three sub-hierarchies that are biological process, molecular function and cellular component. Gene Ontology Annotation@EBI (GOA[4]) [13] provides assignments of GO terms to gene products for all organisms with completely sequenced genome, including human, by a combination of electronic assignment and manual annotation.

The UMLS is made of two major components, the Metathesaurus® (MTH), a repository of 1,179,177 concepts (2005AA release), and the Semantic Network, a limited network of 135 Semantic Types. The MTH is built by merging more than 100 vocabularies, including Medical Subject Headings (MeSH), GO and Genew. In the MTH, synonymous terms are clustered under a same concept, each concept having a unique Concept Unique Identifier (CUI). MTH concepts are related by a set of 22,623,179 relations, including:

- hierarchical relations : 'has parent' and 'has child',
- associative relations named 'other relations', for example - 'has a broader relationship', 'has relationship other than synonymous, narrower, or broader', 'unspecified source asserted relatedness, possibly synonymous', 'the relation is similar or "alike"', 'can by qualified by' or 'source asserted synonymy'-,
- co-occurrences in Medline, with their frequencies. A co-occurrence relation in the MTH corresponds to terms indexing the same article in Medline.

The Semantic Network provides a means to categorize all concepts represented in the MTH. Each MTH concept is assigned to at least one Semantic Type. The 135 Semantic Types can be aggregated into 15 Semantic Groups, e.g. the Semantic Types *Disease or Syndrome* and *Pathologic Function* belong to the Semantic Group *Disorders* [14].

Genew is a database that has been established by the HUGO Gene Nomenclature Committee (HGNC[5]) [15] to address heterogeneity in gene naming and identifiers. For each known human gene, the HGNC has approved a unique gene name and symbol. For a given gene, the Genew database provides this approved nomenclature, as well as various nomenclature information provided by other resources, including Uniprot and Entrez-gene Identifiers.

The process of biomedical annotation via BioMeKE is made of three components : (1) an heterogeneity module that manages the gene naming heterogeneity using Genew, (2) a biological module that provides a biological annotation based on GO and (3) a medical module that provides a medical annotation based on UMLS (Figure 1).

[4] http://www.ebi.ac.uk/GOA/
[5] http://www.gene.ucl.ac.uk/nomenclature/index.html

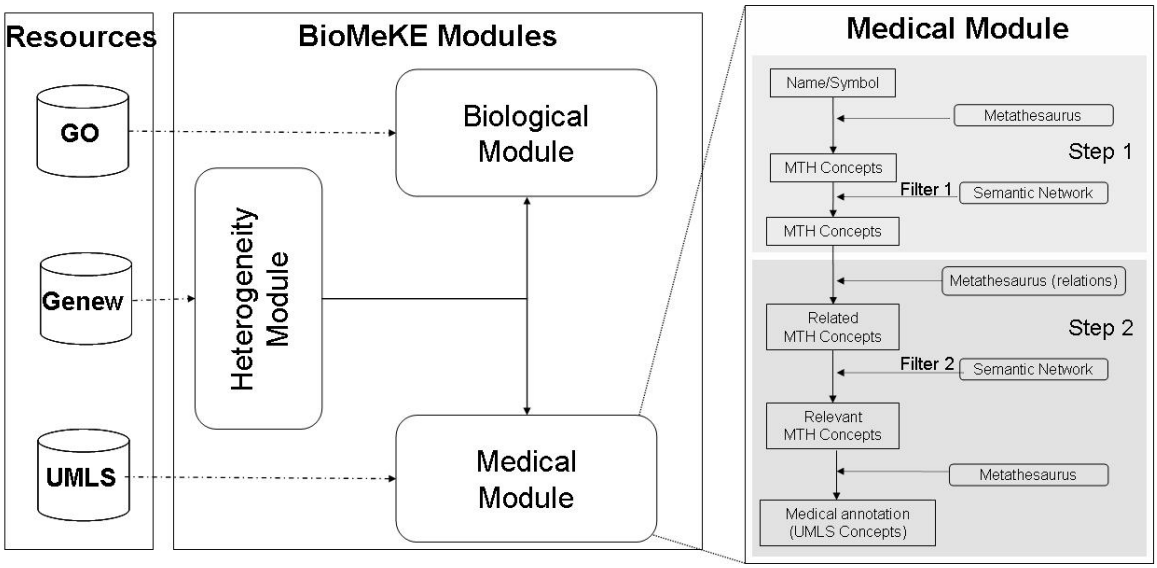

Figure 1 – Architecture and annotation process of BioMeKE.

2.1.1 Heterogeneity module

The biomedical annotation process in BioMeKE starts with the heterogeneity module. For a given gene entry, it uses Genew for providing all its nomenclature elements (name, symbol, aliases) and its identifiers (e.g. Uniprot ID). The identifiers are fundamental to get a wider representation of the gene through existing cross-references in multiple data sources and different name appearances in literature. It is also a prerequisite step to the following annotation modules.

2.1.2 Biological module

The biological module uses the UniProt ID delivered by the heterogeneity module. Based on the UniProt ID, the list of terms associated with a gene product is extracted from GOA. The extracted terms are used as attributes to provide information about the molecular functions, the biological processes and the cellular components related to the gene product.

2.1.3 Medical module

The medical module uses all the nomenclature elements to provide a medical annotation by delivering a list of UMLS concepts associated to the gene. As the UMLS annotation process consists in first finding concepts corresponding to gene in the MTH and second finding its relations in the MTH, it is performed in two steps:

1. *Mapping gene or gene product names to MTH (step 1 in figure 1)*. The objective is to extract the MTH concepts corresponding to a gene. For a given gene, the heterogeneity module provides different nomenclature elements (s.a. name, symbol, aliases). They are successively searched for in the MTH. A filtering phase (Filter 1) is performed to select only the MTH concepts that correspond to the gene, i.e. classified under the five UMLS Semantic Types: - *Gene or Genome*; *Amino Acid, Peptide or Protein; Nucleic Acid, Nucleoside or Nucleotide*; *Molecular Function*; *Disease or Syndrome* –. For example, Ferritin corresponds to two MTH concepts, one of them (C0015879) is assigned to the Semantic Types *Amino Acid, Peptide, or Protein*, the other (C0373607) is assigned to the Semantic Types *Laboratory Procedure*. The latter is not relevant. The Semantic Types *Laboratory Procedure* is absent from the list of relevant Semantic Types (Filter 1). Therefore, C0015879 is selected whereas C0373607 is not.

2. S*earching for MTH concepts to annotate the gene (step 2 in figure 1)*. This step explores MTH relations to perform the medical annotation. For a given MTH concept, the annotation process selects concepts that are related to it through one of the following relations: *parent*, *other relations*, and *co-occurrence*, and assigned to at least one of the 22 relevant Semantic Types (Filter 2) that may be of interest for the interpretation of post genomic data. For example, *Cell or Molecular Dysfunction* belongs to that list whereas *Geographic Area* does not. These 22 Semantic Types are members of seven distinct Semantic Groups. 10 of these Semantic Types are classified under the Semantic Group *Disorders* and 4 under the Semantic Group *Physiology*. An example of a biomedical annotation provided by BioMeKE is illustrated in table 1 with the HFE gene product (UniProt:Q30201).

Table 1 – Biomedical annotation of HFE provided by BioMeKE. The top part of the table presents the HFE nomenclature provided by the heterogeneity module. The left part of the table presents the HFE GO annotations provided by the biological module, grouped under the three GO sub-hierarchies. The right part of the table shows some of the HFE UMLS annotations provided by the medical module and grouped by Semantic Types.

Nomenclature Genew	
Approved Symbol	HFE
Approved Name	hemochromatosis
HGNC ID	4886
Entrez Gene ID	3077
Uniprot ID	Q30201
GO annotations	**UMLS annotations**
Molecular Function	***Genetic Function***
MHC class I receptor activity	Genetic Markers
Biological Process	Multifactorial Inheritance
protein complex assembly	***Neoplastic Process***
transport	Bile Duct Neoplasms
iron ion transport	Cholangiocarcinoma
iron ion homeostasis	Liver neoplasms
receptor mediated endocytosis	Primary carcinoma of the liver cells
immune response	***Organ or Tissue Function***
antigen presentation, endogenous antigen	Intestinal Absorption
antigen processing, endogenous antigen via MHC class I	***Pathologic Function***
Cellular Component	Insulin Resistance
cytoplasm	Tachycardia, Ventricular
integral to plasma membrane	Hypertrophy, Right Ventricular
	Hyperpigmentation

BioMeKE is implemented as a Java Swing application that relies on JTree, JTable and other GUI components. We have wrapped BioMeKE as a Java Web Start application which provides the advantage to check before any download if a new version of the application is available. BioMeKE is freely available at http://www.med.univ-rennes1.fr/biomeke/.

2.2 Data mining of enriched transcriptomic data

Data mining of enriched transcriptomic data is performed in the data warehouse GEDAW (Gene Expression DAta Warehouse) [12].

2.2.1 Data warehouse architecture and schema

GEDAW is an object oriented data warehouse devoted to transcriptomic data analysis. GEDAW schema includes three data domains : 1) experimental division, i.e. gene expression measurements through several physiopathological conditions 2) genomic division, i.e. gene,

mRNA, protein sequences and their annotations and 3) biomedical knowledge, i.e. biological and medical concepts that annotate the genes.

Data sources used for the integration process are local or spread world wide and hosted on different representation systems, each having its own schema.

A local relational database is used to populate the experimental domain of the warehouse. It is a MIAME [16] compliant database locally built as a repository of array data storing as many details as possible on methods used, the protocols and the results obtained.

XML records from GenBank[6] [17] have been used to instantiate the genomic domain of GEDAW.

GO and UMLS concepts delivered by BioMeKE as a XML document are used to integrate the biomedical knowledge.

A unique global schema has been designed to conciliate experimental, genomic and biomedical genes information. Java is used for the description and the instantiation of the classes. The ODBMS (Object DataBase Management System) Versant FastObjects[7] is used to make the Java Objects persistent.

We have developed an automatic integration process through the use of mapping. Through mapping rules at the schema level, elements and concepts of GenBank, GO and UMLS are selected, extracted and integrated. Through mapping rules at the instance level, problems of heterogeneity in gene identification occurring in GEDAW are resolved. In fact, in GEDAW, to several GenBank accession numbers can correspond a same gene product. These rules (at the instance level) use the output of the heterogeneity module provided by BioMeKE that corresponds to the full Genew nomenclature associated to a gene. Approved symbols and names are used to identify the identical GenBank identifiers in GEDAW.

2.2.2 Data analysis procedure

With the overall integrated knowledge, the warehouse provides an analysis environment where experimental data can be mined through workflows that combine successive analysis steps.

GEDAW supports several functions for microarray data analysis, consisting of either internal or external analyses applied to the group of genes of interest – these genes resulting from a database selection query according to one or more criteria. Internal analyses retrieve information about the selected genes thanks to APIs that use OQL (Object Query Language) and Java. External analyses use external bioinformatics tools applied to integrated data. These two kinds of analyses may be combined to create successive steps, thus forming a workflow.

Many specific workflows have been designed in the context of microarray analysis. One of them has been designed according to the hypothesis that genes sharing an expression pattern should be associated. It has been used in order to find out new genes associated to a disease.

More specifically, the strategy consists in selecting a group of genes that are associated with the same disease and a typical expression pattern, and then extrapolating this group to more genes involved in the disease by searching for expression pattern similarity. The genes are then characterized by the corresponding biological processes and cellular components using integrated GO annotations. The strategy is divided in four steps:

---

[6] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
[7] http://www.versant.com/

- *Selection of objects*: A group of genes sharing the same UMLS annotation is selected, by querying the integrated objects in GEDAW using OQL and Java.

- *Visualization*: The obtained gene names and associated expression ratios are then visualized. This is done by searching for genes attributes in the warehouse with specific queries using OQL and Java.

- *External analyses*: Gene expression ratios are then analysed using the J-Express Pro software package[8] (2.7 version) [18]. The K-Means clustering analysis method, applied to the group of genes, provides clusters of genes presenting different expression patterns. The Closest Neighbours analysis method is then performed to identify the genes represented on the microarray that have similar patterns to those obtained by K-Means clustering. Genes found by the Closest Neighbours analysis extend the initial clusters.

- *Internal analysis*: the genes of these extended clusters are then characterized, by searching for the most represented GO biological processes. This is performed by specific OQL queries on the GO terms integrated in GEDAW.

3 Application

3.1 Data set

The workflow described above is used in combination with BioMeKE to identify new genes that could be associated to liver diseases and to characterize their expression patterns and the biological processes in which they are involved.

Liver diseases, including those from infectious, alcoholic, metabolic, toxic and vascular etiologies, are a major public health problem [19]. Indeed, they are frequently complicated by the occurrence of liver failure or the development of cirrhosis or liver cancer. Despite such a strong impact, molecular mechanisms involved in the occurrence of these diseases and of their complications are not fully understood. Therefore, studies are conducted in order to identify new molecular mechanisms, and thus to develop new diagnostic and therapeutic tools which will allow a better care of patients.

In this study, we used a human liver devoted cDNA microarray on which 2472 cDNAs are deposited and we studied gene expression modulation during the hepatic HepaRG cell line differentiation process [20]. This human cell line has the originality, under controlled culture condition, to evolve from a bipotent proliferative population towards both differentiated hepatocyte-like and biliary-like cells [21, 22]. Therefore, HepaRG cell line is a valuable model for studying the shift between differentiated functional hepatocytes and biliary cells to altered proliferative cells, as observed in some liver diseases.

The integration process was performed to store all the annotations of the genes spotted on the microarray. The data unification process, using gene nomenclature identified 584 distinct genes on the 2472 deposited cDNAs. We then used the analysis workflow, described in figure 2, to find and characterize genes associated to liver diseases (figure 2). More specifically, we focused on studying the genes known to be associated to liver diseases and relating their expression patterns to genes of the array.

[8] http://www.molmine.com/frameset/frm_jexpress.htm

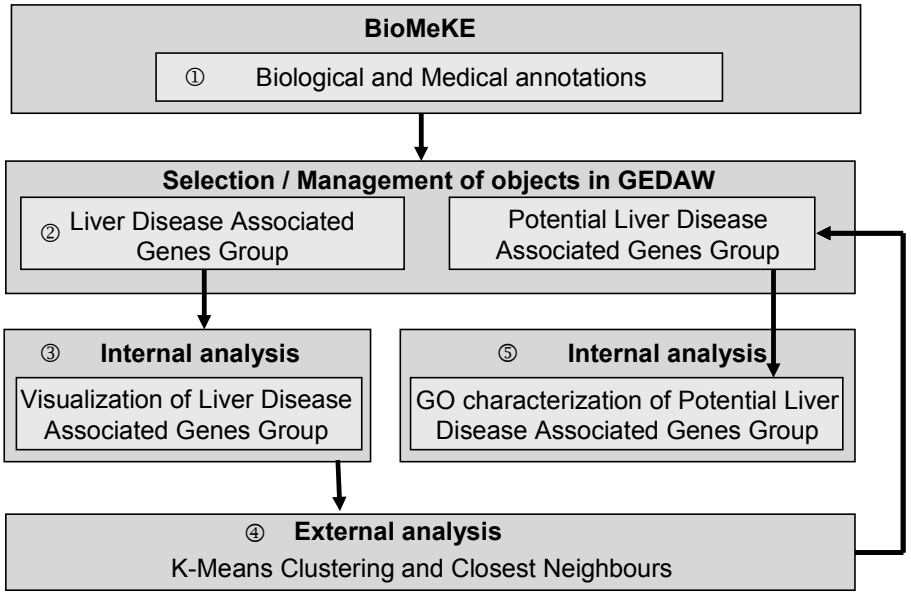

Figure 2 –Analysis Workflow for liver disease group. It is divided in 5 successive tasks: 1) The genes are annotated thought BioMeKE, 2) Selection of genes that are annotated by liver disease terms, they constitute the Liver Disease Associated Genes Group, 3) Data about the Liver Disease are visualized, 4) K-Means and then Closest neighbours algorithms are applied creating a new group: the Potential Liver Disease Associated Genes Group, 5) The genes of the Potential Liver Disease Associated Genes Group are characterized by a GO analysis to find the biological processes and the cellular components mostly represented.

3.2 Results

Here, we present the results that we have obtained for the data set presented in section 3.1.

1. Biomedical annotation through BioMeKE - BioMeKE provided GO annotation for 437 (74.8%) genes. 381 (65.2%) genes have been found in the UMLS. Among these 381 genes, 173 (45.4%) have relations in the UMLS (i.e. have UMLS annotation). The UMLS annotation corresponds to three types of relations in the MTH: i) 85 genes have annotations under the relation *parent*, ii) 129 genes have annotations under the relation *other relations* and iii) 33 genes have annotations under the relation *co-occurrence*. For example, the gene Beta-2 microglobulin has been annotated by 'Alpha-Globulins' as *parent* relation, by 'incomplete anencephaly, hemicrania' as *other relations* and by 'Hepatitis B, Chronic' as *co-occurrence* relation. Most of the genes have annotations under the Semantic Group *Chemical and Drug*. Among the 173 annotated genes, 42 genes have annotations under the Semantic Groups *Disorders* and *Physiology*.

2. Selection of objects in GEDAW: creation of a Liver Disease Associated Genes Group – Genes of the array that are annotated by "liver disease" and their descendants in the UMLS are selected. This group is called Liver Disease Associated Genes Group.

3. Visualization of the Liver Disease Associated Genes Group – Characteristics of the genes that belong to the Liver Disease Associated Genes Group are visualized by the user, including the gene name. We found 42 concepts corresponding to liver diseases including liver cirrhosis (CUI:C0023890), hepatitis B (CUI:C0524909) or hemochromatosis (CUI:C0018995) and 18 genes annotated by at least one of those 42 concepts (see Table 2).

Table 2– Genes annotated by at least one child concept of "liver disease" concept. The table shows a part of the nomenclature of the genes that are annotated by at least one child concept of "liver disease" concept.

Approved Symbol	Approved Name	RefSeq ID	Entrez Gene ID
AFP	alpha-fetoprotein	NM_001134	174
ALB	albumin	NM_000477	213
B2M	beta-2-microglobulin	NM_004048	567
C16orf5	chromosome 16 open reading frame 5	NM_013399	29965
CAT	catalase	NM_001752	847
CFHL5	complement factor H-related 5	NM_030787	81494
CXCR4	chemokine (C-X-C motif) receptor 4	NM_003467	7852
CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1	NM_000773	1571
EPO	erythropoietin	NM_000799	2056
FN1	fibronectin 1	NM_212482	2335
FNDC3A	fibronectin type III domain containing 3A	NM_014923	22862
G6PD	glucose-6-phosphate dehydrogenase	NM_000402	2539
HFE	hemochromatosis	NM_139011	3077
HP	haptoglobin	NM_005143	3240
PIK3AP1	phosphoinositide-3-kinase adaptor protein 1	NM_152309	118788
PLG	plasminogen	NM_000301	5340
TF	transferrin	NM_001063	7018
TGFA	transforming growth factor, alpha	NM_003236	7039

4. External analyses: K-Means and Closest Neighbours – Two successive external analyses are performed on the set of genes that belong to the Liver Disease Associated Genes Group in order to propose new genes associated to liver diseases: K-Means clustering analysis of their expressions and then Closest neighbours analysis. In the context of HepaRG differentiation experiments, four differentiation stages have been studied through six comparisons (Figure 3). Therefore six expression ratios per gene have been delivered to the K-Means program, for the 14 genes among the 18 for which we had valid and normalized expression ratios.

Four distinct patterns have been found by the K-Means analysis. The first pattern is that of AFP, EPO and G6PD, the second one is that of CAT, CXCR4, CYP2E1 and PLG, the third one is that of ALB, HP and TF and the last one is that of the remaining genes of the Known Liver Disease Marker Group: B2M, C16orf5, FN1 and HFE. The Closest Neighbours analysis created four gene clusters associating genes of the array that have similar patterns of those found by the K-Means clustering (Figure 4).

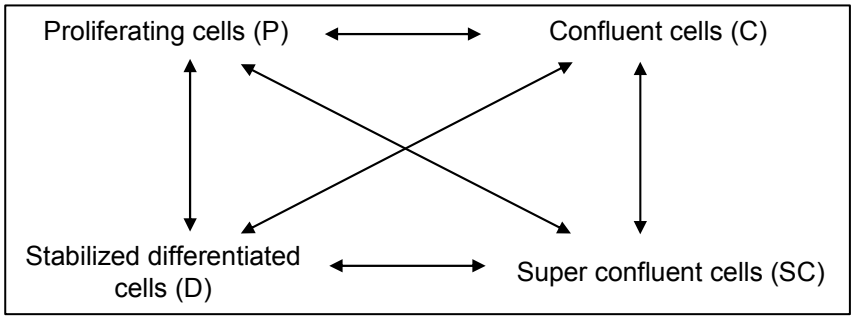

Figure 3 – Experimental design of HepaRG differentiation hybridizations. HepaRG differentiation process is studied through four stages: (P) proliferating cells, 3 days post-spreading, (C) confluent cells, 5-6 days post-spreading, (SC) super confluent cells, 12-15 days post-spreading and finally (D) stabilized differentiated cells, 30 days post-spreading with the last 15 days in basal medium supplemented with 2% of DMSO (dimethyl sulfoxyde). The six comparisons that have been made for the study are represented by the arrows.

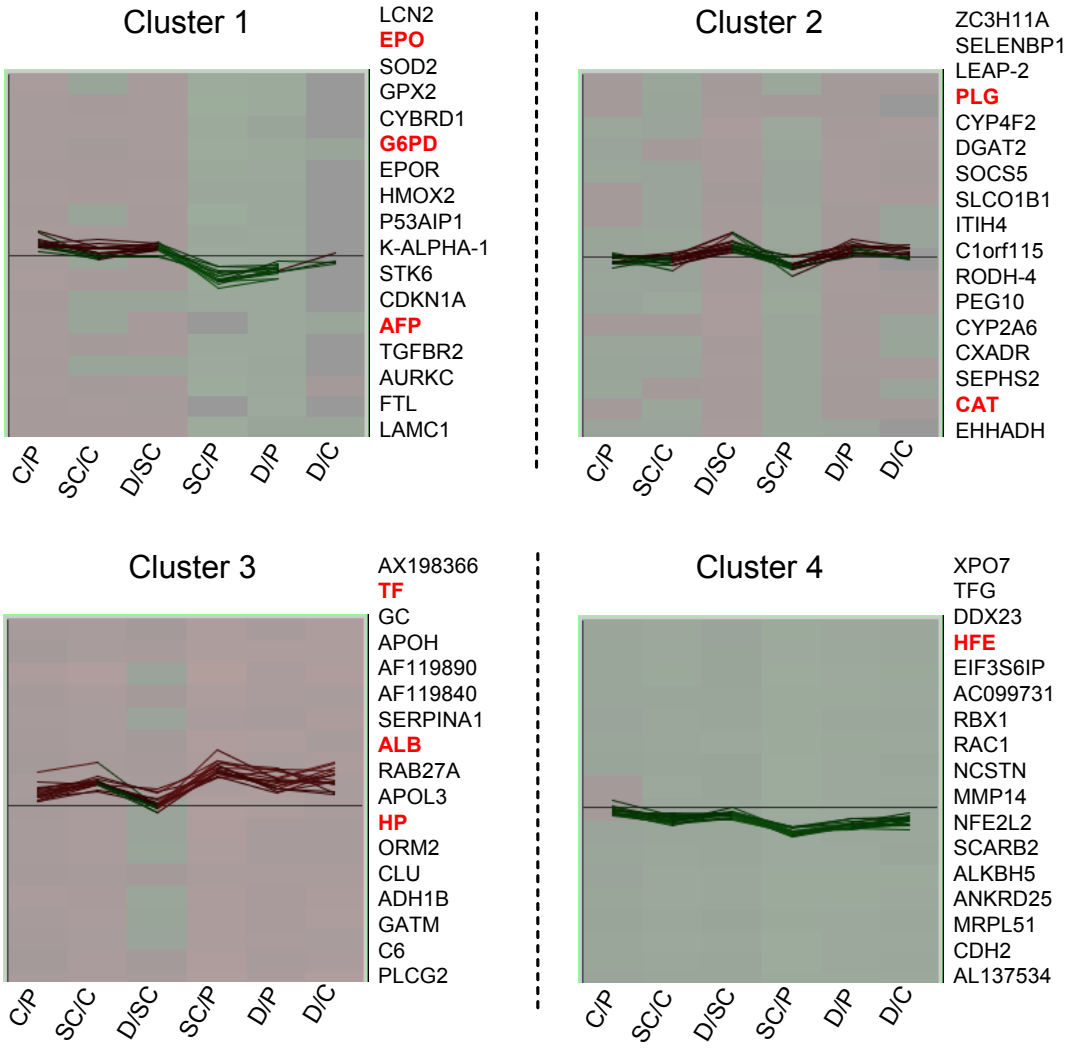

Figure 4 – Pattern characterization of the Potential Liver Disease Associated Genes Group. For each cluster, the expression pattern is represented with the list of genes associated to the cluster. The patterns are composed of six points that correspond to the six comparisons of the HepaRG differentiation study. The genes are defined with their respective HGNC approved symbols. Red symbols correspond to the genes that belong to the Liver Disease Associated Genes Group, whereas the black symbols correspond to genes of the Potential Liver Disease Associated Genes Group that have been identified by the Closest Neighbours analysis.

Each cluster contains 17 genes. Cluster 1 was created from the pattern of AFP, EPO and G6PD, cluster 2 from that of CAT, CXCR4, CYP2E1 and PLG, cluster 3 from ALB, HP and TF, and cluster 4 from B2M, C16orf5, FN1 and HFE. The patterns of clusters 1 and 2 correspond to genes highly expressed during the late stage of differentiation (D/SC). The patterns of the clusters 1 and 2 are different in the last comparison that is made between stabilized differentiated cells and proliferating cells (D/P) (Figures 3 and 4). The pattern of cluster 3 corresponds to genes highly expressed in the early stage of differentiation (SC/C). The pattern of cluster 4 corresponds to genes under expressed during the whole process of differentiation.

The genes found in the four clusters are considered as potential genes of interest during liver diseases, and belong to a new group called the Potential Liver Disease Associated Genes Group. Starting from the 14 genes known to be involved in liver metabolism (red symbols in table 3), this new group represent a set of 59 genes of interest (black symbols in table 3). Some of those 59 genes are known by the experts to be involved in liver metabolism, such as

the apolipoprotein H (APOH in cluster 3) [23], the alcohol deshydrogenase (ADH1B in cluster 3) [24] and the cytochromes (CYP4F2 and CYP2A6 in cluster 2) [25]. However, some are not clearly associated to hepatic function, such as the apolipoprotein L3 (APOL3 in cluster 3) [26] and some have not yet been described [GenBank: AF119890, AF119840 and AL137534, corresponding to mRNA sequences; AX198366 and AC099731 corresponding to DNA sequence].

5. Internal analysis: Gene Ontology characterization of Potential Liver Disease Associated Genes Group – We studied the GO biological processes and the GO cellular components represented in these four clusters of genes belonging to the Potential Liver Disease Associated Genes Group, to characterize the genes. The results are presented in Table 3.

Table 3 - Biological characterization of the four clusters of Potential Liver Disease Associated Genes Group. The biological characterization of the four clusters has been performed using Gene Ontology. The results concerning the frequency of annotated genes per the five most frequent Biological processes are represented in A. The results concerning the frequency of annotated genes per the three most frequent Cellular components are represented in B. The same color code has been used for tables A and B: red corresponds to - more than 66% of genes - , green corresponds to - less than 33% of genes -, and white corresponds to - between 33 and 66% of genes.

A

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Signal transduction				
Transport				
Cellular metabolism				
Response to stimulus				
Regulation of cellular process				

B

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Intracellular				
Membrane				
Extracellular region				

Five GO biological processes are frequently represented among the four clusters: *signal transduction* [GO:0007165], *transport* [GO:0006810], *cellular metabolism [GO:0044237], response to stimulus* [GO:0050896], *regulation of cellular process* [GO:0048522]; and three GO cellular components: *intracellular* [GO:0005622], *membrane* [GO:0016020] and *extracellular region* [GO:0005576].

The proportions of genes per cluster annotated by these terms have been calculated. In cluster 1, 2 and 4 the biological process mostly represented is *cellular metabolism*; in cluster 3, the over-represented biological process is *response to stimulus*. In cluster 3 the cellular component mostly represented is *extracellular region*. In cluster 4, the over-represented cellular component is *intracellular*. There is no over or under-represented cellular component in cluster 1 and 2.

Therefore, it seems that the clusters 1, 2 and 4 are mainly composed of genes involved in cellular metabolism whereas the cluster 3 is mainly composed of genes involved in immune response and coding for secreted products. This results show that genes having different expression patterns or biological processes involvements could be associated with liver diseases. This suggests that the mechanisms of involvement of these genes in liver diseases and their ways of action are different. These genes need to be biologically investigated to have a better understanding of their involvement in liver diseases.

4 Discussion and conclusion

This paper has presented our experience in combining experimental and genomic data with biomedical knowledge, using BioMeKE a biomedical annotation system and GEDAW a gene expression data warehouse. This has been applied to extract relevant knowledge from liver microarray experiments.

BioMeKE is a system whose originality is to be based on a medical vocabulary, the UMLS. If most of the systems use Gene Ontology to annotate sets of genes with information on their functions, few similar efforts have been made to provide clinical and medical information on genes. Like BioMeKE, GenesTrace™ [27] uses the UMLS but provides a list of genes related to a disease whereas our system provides medical concepts associated to a gene. GenesTrace uses the relationship existing between UMLS diseases and other UMLS concepts, restricted to GO. Among the 200,000 MTH diseases concepts, they found 1,407 diseases concepts that are associated with at least one GO term, and among these 1,407, they found 142 distinct genes that are related to their specific disease concept.

In addition, issues related to heterogeneity in gene nomenclature among different sources are addressed in BioMeKE. By integrating not only the GO and the UMLS "ontologies" but also the Genew resource, BioMeKE is able to provide a biomedical annotation associated to a gene as well as the full information about its nomenclature. This aspect has been particularly primordial for the data conciliation in GEDAW: approved HGNC symbols and names have been used to unify the various GenBank identifiers corresponding to a same gene product. BioMeKE provides more cleansed, conciliated and non redundant data to mine.

However, BioMeKE does not extract information from OMIM [28], which associates mutations of one given gene with the corresponding genetic diseases. Indeed, OMIM provides textual description of the genetic diseases related to gene mutations and BioMeKE does not include natural language processing nor text mining modules. However, OMIM terminology is integrated in the UMLS. Therefore, some annotating UMLS concepts correspond to OMIM terms, e.g. an annotating UMLS concept for Transferrin, is 'Insulin-Like Growth Factor II', whose sources include OMIM terminology.

Since the 2004AA version of the UMLS, GO has been part of the MTH [29, 30]. This merging is based on exact matching and normalized matching. During the merging of GO in the UMLS, they have shown that 23.03% of the GO terms 'match' with a concept that is represented by another source vocabulary. Among the 23.03% of these concepts, 19.74% correspond to the MeSH vocabulary and 11.05% correspond to SNOMED [31].

Biomedical annotation provided by BioMeKE is based on co-occurrence relations among the three above cited groups of relations (cf. section 2.2.1). Concepts related by a co-occurrence relation in the MTH correspond to terms indexing the same article in Medline. For each pairs of MeSH descriptors, the frequency of co-occurrence in Medline citations is recorded in the UMLS. In contrast to the relationships asserted within source vocabularies, the co-occurrence relationships in the MTH can connect very different concepts, such as genes and diseases. In further development of BioMeKE, text mining methods could be added to extract genes related to biomedical concepts.

Even if GO is merged in the UMLS, it is important to keep two different annotation processes. In fact, among the 173 annotated genes by UMLS, only 17 genes are annotated by UMLS concepts whose source is GO (UMLS-GO annotation), and only 7 genes exhibit redundant annotations, i.e. equivalent UMLS-GO annotations and GO annotations (through GOA). The major limitation of BioMeKE is that only few genes are related to biomedical concepts in the UMLS. Nevertheless, by using both GO and UMLS, BioMeKE provides complementary and valuable biological and medical information on genes.

The strength of our approach is to combine biomedical concepts with gene expression data enriched with genomic data from GenBank in an environment where complementary data are conciliated and locally available for retrieval and analysis. Thereafter, efficient analyses on experimental data can be done, taking advantage of the integrated biomedical knowledge through workflows of successive internal and external analyses.

Nevertheless, lessons have been learned during the development of GEDAW and current works are ongoing on quality issues of the biomedical data before their integration and mining: duplicates, errors, contradictions, inconsistencies for correcting and ensuring information quality when data come from different sources with different degrees of quality and trust.

The work presented in this paper has also demonstrated the finality of the warehousing approach applied to bioinformatics: bio-data integration, supervised analyses, and knowledge extraction. Being conscious that analysis requirements evolve with constant emergence on the Web of new complex data types like protein structures, gene interactions or metabolic pathways, workflows in GEDAW are evolving as well.

The effectiveness of our combined approach has been evaluated in the context of liver transcriptome study. Starting from a group of genes annotated in GEDAW by UMLS terms associated to liver disease, we have been able to identify new genes potentially associated to occurrence and/or development of liver diseases. Some of those genes were known to be associated to liver metabolism, whereas some others not. They have been biologically characterized and are associated to different biological processes. Their impact in biological pathways as well as their use as biological markers or therapeutic targets remain to be evaluated. This work will be conducted by researchers using molecular biology techniques, including gene expression study in physiopathological conditions in patients and in animal models.

Acknowledgements

This work was supported by grants from Region Bretagne (20046805, and PRIR139) and MRT fellowship (EG).

5 References

[1] P. Anderle, M. Duval, S. Draghici, A. Kuklin, T. G. Littlejohn, J. F. Medrano, D. Vilanova, and M. A. Roberts, "Gene expression databases and data mining," *Biotechniques*, vol. Suppl, pp. 36-44, 2003.

[2] G. Piatetsky-Shapiro and P. Tamayo, "Microarray Data Mining : Facing the Challenges," in *ACM SIGKDD, Explorations*, vol. 5, S. Sarawagi, Ed., 2003, pp. 1-5.

[3] S. Davidson, C. Overton, and P. Buneman, "Challenges in integrating biological data sources," *Journal of Computational Biology*, vol. 2, pp. 557-572, 1995.

[4] T. Hernandez and S. Kambhampati, "Integration of biological sources: current systems and challenges ahead," *SIGMOD record*, vol. 33, pp. 51-60, 2004.

[5] N. W. Alkharouf, D. C. Jamison, and B. F. Matthews, "Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases," *J Biomed Biotechnol*, vol. 2005, pp. 181-8, 2005.

[6] M. Cornell, N. W. Paton, C. Hedeler, P. Kirby, D. Delneri, A. Hayes, and S. G. Oliver, "GIMS: an integrated data storage and analysis environment for genomic and functional data," *Yeast*, vol. 20, pp. 1291-306, 2003.

[7] K. Fellenberg, N. C. Hauser, B. Brors, J. D. Hoheisel, and M. Vingron, "Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis," *Bioinformatics*, vol. 18, pp. 423-33, 2002.

[8] H.-H. Do and E. Rahm, "Flexible Integration of Molecular-Biological Annotation Data: The GenMapper Approach," presented at 9th International Conference on Extending Database Technology, Heraklion, Crete, Greece, 2004.

[9] T. Kirsten, H. H. Do, and E. Rahm, "A Data Warehouse for Multidimensional Gene Expression Analysis," Working Paper, University of Leipzig, Working Paper November 2004 2004.

[10] The Gene Ontology Consortium, "Gene ontology: tool for the unification of biology," *Nat Genet*, vol. 25, pp. 25-9, 2000.

[11] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res*, vol. 32, pp. 267-70, 2004.

[12] E. Guérin, G. Marquet, A. Burgun, O. Loréal, L. Berti-Equille, U. Leser, and F. Moussouni, "Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW," presented at 2nd International Conference on Data Integration in Life Sciences, San Diego, California, USA, 2005.

[13] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Res*, vol. 32, pp. 262-6, 2004.

[14] A. T. McCray, A. Burgun, and O. Bodenreider, "Aggregating UMLS semantic types for reducing conceptual complexity," *Medinfo*, vol. 10, pp. 216-20, 2001.

[15] H. M. Wain, M. J. Lush, F. Ducluzeau, V. K. Khodiyar, and S. Povey, "Genew: the Human Gene Nomenclature Database, 2004 updates," *Nucleic Acids Res*, vol. 32, pp. 255-7, 2004.

[16] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nat Genet*, vol. 29, pp. 365-71, 2001.

[17] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Res*, vol. 33, pp. 34-8, 2005.

[18] B. Dysvik and I. Jonassen, "J-Express: exploring gene expression data using Java," *Bioinformatics*, vol. 17, pp. 369-70, 2001.

[19] S. Sherlock and J. Dolley, *Diseases of the liver and biliary system*, 11 ed: Blackwell Science, 2002.

[20] M. B. Troadec, D. Glaise, G. Lamirault, M. Le Cunff, E. Guerin, N. Le Meur, L. Detivaud, P. Zindy, P. Leroyer, I. Guisle, H. Duval, P. Gripon, N. Theret, K. Boudjema, C. Guguen-Guillouzo, P. Brissot, J. J. Leger, and O. Loreal, "Hepatocyte iron loading capacity is associated with differentiation and repression of motility in the HepaRG cell line," *Genomics*, vol. 87, pp. 93-103, 2006.

[21] P. Gripon, S. Rumin, S. Urban, J. Le Seyec, D. Glaise, I. Cannie, C. Guyomard, J. Lucas, C. Trepo, and C. Guguen-Guillouzo, "Infection of a human hepatoma cell line by hepatitis B virus," *Proc Natl Acad Sci U S A*, vol. 99, pp. 15655-60, 2002.

[22] R. Parent, M. J. Marion, L. Furio, C. Trepo, and M. A. Petit, "Origin and characterization of a human bipotent liver progenitor cell line," *Gastroenterology*, vol. 126, pp. 1147-56, 2004.

[23] A. Steinkasserer, D. J. Cockburn, D. M. Black, Y. Boyd, E. Solomon, and R. B. Sim, "Assignment of apolipoprotein H (APOH: beta-2-glycoprotein I) to human chromosome 17q23----qter; determination of the major expression site," *Cytogenet Cell Genet*, vol. 60, pp. 31-3, 1992.

[24] D. W. Crabb, M. Matsumoto, D. Chang, and M. You, "Overview of the role of alcohol dehydrogenase and aldehyde dehydrogenase and their variants in the genesis of alcohol-related pathology," *Proc Nutr Soc*, vol. 63, pp. 49-63, 2004.

[25] J. P. Villeneuve and V. Pichette, "Cytochrome P450 and liver diseases," *Curr Drug Metab*, vol. 5, pp. 273-82, 2004.

[26] N. M. Page, D. J. Butlin, K. Lomthaisong, and P. J. Lowry, "The human apolipoprotein L gene cluster: identification, classification, and sites of distribution," *Genomics*, vol. 74, pp. 71-8, 2001.

[27] M. N. Cantor, I. N. Sarkar, O. Bodenreider, and Y. A. Lussier, "Genestrace: phenomic knowledge discovery via structured terminology," *Pac Symp Biocomput*, pp. 103-14, 2005.

[28] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res*, vol. 33, pp. D514-7, 2005.

[29] O. Bodenreider, J. A. Mitchell, and A. T. McCray, "Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics," *Proc AMIA Symp*, pp. 61-5, 2002.

[30] J. Lomax and A. Mc Cray, "Mapping the Gene Ontology into the unified medical language system," *Comp Funct Genomics*, vol. 5, pp. 345-361, 2004.

[31] J. Lomax, "The Gene Ontology and its insertion into UMLS," presented at Standards and Ontology for Functional Genomics (SOFG), 2002.