The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

BioMeKe : an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis

Gwenaëlle Marquet [1], Anita Burgun [1], Fouzia Moussouni[2], Emilie Guérin [2], Franck Le Duff [1], Olivier Loréal [2]

[1] *Laboratoire d'Informatique Médicale – Faculté de Médecine – 35043 RENNES Cedex - France*
[2] *INSERM U522, CHU Pontchaillou, 35033 RENNES Cedex, France*

{gwenaelle.marquet, anita.burgun}@univ-rennes1.fr

Abstract

Semantic interoperability between knowledge bases in medicine, and knowledge base in genomics and molecular biology will lead to advances in fundamental research as well as to improved patient care . DNA chips strategy is used for transcriptome analysis in order to identify deregulated genes in physio-pathological conditions. The objective of the BioMedical Knowledge Extraction project (BioMeKe) is to develop a knowledge warehouse in the context of transcriptome analysis during liver diseases. Knowledge sources include ontologies, related terminologies and annotations linked towards public databases (e.g., SWISSPROT). BioMeKe has been developed to have access to information using systematic investigation upon a concept, gene, gene products, pathology, or any target keyword, and is based on the combination of several relevant resources: UMLS, GeneOntology, MeSH supplementary terms, GOA, and HUGO. Current efforts are focusing on exploiting this ontology-based Knowledge Extractor, to enrich the expression data on genes delivered by a liver specific DNA microarray for better assistance of analysis.

Keywords:
Knowledge extraction, Transcriptome, Ontology, UMLS, GeneOntology, microarray

1. Introduction

In genomics, the microarray technology provides experimental data on gene expression profiles [1,2], that are then stored in data warehouses [3]. For exploiting these data, biologists need first to have access to all the knowledge available on these genes, in order to be able to interpret the experiments and generate new physio-pathologic hypothesis. However the knowledge is kept in independent heterogeneous databases, that offer limited interoperability [4]. The paper presents BioMeKe[1], a Biomedical Knowledge Extraction system, applied to liver specific DNA-chips. The presented prototype was designed to search relevant information in existing knowledge databases of two types : ontologies and public databanks. The ontologies include Unified Medical Language System® covering the whole biomedical domain, and GeneOntology™ focusing on genomics. These ontologies are integrated in BioMeke, while public databases can only be accessed through Internet queries. In the final version, the genes warehouse currently in progress will store experimental data (from DNA chips), as well as the extracted knowledge. BioMeKe will be acting as a server automatically extracting knowledge from ontologies via the Internet. It

[1] http://www.med.univ-rennes1.fr/~marquet/biomeke.html

will also be associated with Gene Expression Data Warehouse (Gedaw), developped for transcriptome analysis in the domain of liver disease [3], presented in [5].

2. Material and Methods

Materials

It consists of the selected ontologies and public databanks.

UMLS:

The Unified Medical Language System® (UMLS®)[2] has been developed by the US National Library of Medicine since 1986 [6][7]. It is intended to help health professionals and researchers use biomedical information from different sources and is made by mapping many existing terminologies within a unifying framework. It comprises three knowledge bases: the Metathesaurus®, a large repository of concepts, the Semantic Network, a limited network of 134 semantic types, and the Specialist Lexicon which corresponds to lexical resources. The 2002AB edition of the Metathesaurus includes 776,940 concepts and approximately 11,137,725 relationships. The UMLS is intended to cover the whole medical domain. Several projects have mentioned the UMLS with application to the clinical genetics and molecular biology, e.g., [8,9,10,11].

GeneOntology:

GeneOntology™ (GO)[3] is an ontology for molecular biology and genomics developed by the European Consortium at EBI (European Biological Institute) [12]. It is organized with three top categories:

- Molecular Function: a task performed by individual gene products (e.g., transcription factor, DNA helicase)
- Biological Process: a biological goal accomplished via one or more ordered assemblies of molecular functions (e.g., cell growth and maintenance, cAMP biosynthesis)
- Cellular Component: a subcellular structure, location, or macromolecular complex (e.g., nucleus, telomere, origin recognition complex)

As of July 2002, ignoring concepts marked as obsolete in the database, GO contains 5017 process, 4992 molecular function and 1035 component concepts (called terms in GO). Definitions of terms come from the Oxford Dictionary of Molecular Biology and SWISSPROT [13] A gene product has one or more molecular functions and is used in one or more biological processes; it may be associated with one or more cellular components. GO itself is not populated with gene products. GO concepts are to be used as attributes of gene products by collaborating external databases, which can make database cross-references between GO concepts and objects in their database (typically, gene products, or their surrogates, genes). Among the gene product databases, GO Annotation @EBI (GOA), Compugen, Gene Ontology Association Data, and SWISSPROT contribute to assignments of gene products to the GO resource. For each term, they provide links towards molecular function (implicitly has-function), biological process (implicitly has-process), and cellular component (implicitly has-location) in GO.

[2] http://www.nlm.nih.gov/research/umls/

[3] http://www.geneontology.org

GeneOntology Annotation:

GeneOntology Annotation (GOA) is developed by EBI[4]. The objective is to join gene products with GO terms. Gene products that are represented in GOA files come from a public database. To each gene product is assigned an accession number, which provides a means to access the corresponding record in the databanks and bibliography.

MeSH Supplementary Terms public:

In addition to the MeSH core thesaurus, supplementary terms exist in MeSH in order to cover the entire field of molecular biology. In the 2002 release of MeSH in ASCII format, a specific MeSH file contains 129972 records corresponding to chemicals, genes and gene products. Available information in MeSH Supplementary records include name (N1), name of substance (NM), synonyms (SY).

HUGO:

In experimental molecular biology recorded in public resources and bibliography, a gene can have different names. When a biologist searches for information about a given gene, he is not sure to find all the information when the query is performed using a single name.

HUGO[5] provides gene names and their synonyms as well as their symbols [14]. In HUGO database, we can find HUGO number, RefSeq (the NCBI Reference Sequence Project RefSeq provides reference sequence standards for the naturally occurring molecules of the central dogma, from chromosome to mRNAs to proteins), SWISSPROT or LocusLink accession number, the official symbol and the gene name.

Public databanks:

Public databanks, e.g., SWISSPROT[6], GenBank[7], LocusLink[8] provide supplementary information on genes, sequences and proteins, discovered upon a published experiment.

Methods

The first step was to identify the knowledge databases that BioMeke would access, this phase having been done in concertation with biologist. The second step was to understand the way these databases would be accessed, and their possible interaction. The software was then chosen. The ontologies will be stored in a relational SGBD, and the programming language will be object-oriented in order to ease the access to the Gedaw base. The choice was made on MySQL and Java, both available freely on the World Wide Web and portable to Windows and Linux plateforms. Finally the prototype was developped and validated.

3. Results

BioMeke has two main functionnalities : extracting knowledge from integrated ontologies, and presenting results of Internet queries of public databanks.

Knowledge extraction from terminologies and ontologies

The search process may be initiated by inputing to BioMeKe interface either a general term (e.g. « iron metabolism ») or a gene product (e.g. «ferritin heavy chain»). If the user types a term, he/she can choose to search for it either in the UMLS or in GeneOntology. The following procedure is performed:

- Searching in UMLS.

[4] http://www.ebi.ac.uk/

[5] http://www.gene.ucl.ac.uk/nomenclature/

[6] http://www.expasy.org/sprot/

[7] http://www.ncbi..nkm.nih.gov/Genbank/

[8] http://www.ncbi.nlm.nih.gov/LocusLink/

If the term is found in the UMLS, the context of this concept in the UMLS is displayed (categorisation according to the Semantic Network, relationships with other concepts, co-occurrences). If the term is not found in the UMLS, the system will search for it automatically in the MeSH supplementary terms (MeSH-ST) vocabulary. The user can then choose a term from the information extracted from MeSH-ST, and, if present in the UMLS, can visualise its context.

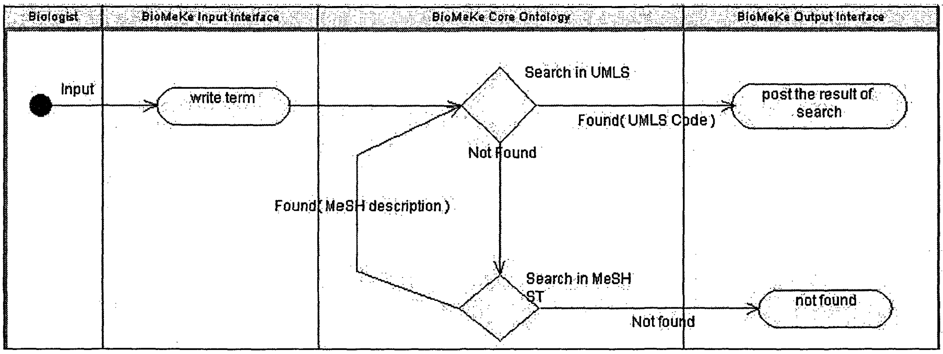


Fig 1 : Search in UMLS

- Searching in GO.

If the term is found in GO, the user can visualise its context in GO: type, relationships, and its context in GOA (i.e. assignments and bibliographic references). If the term is not found in GO, BioMeKe will search for it in GOA, then the user visualises its context in GOA and can follow links to GO through annotations. Figure 3 presents a screen view of the results for the term «iron homeostasis» .

The biologist can also mention explicitly that the entry term is an accession number, a symbol, or the name of a given protein or gene. In this case, BioMeKe searches for the item in GOA. If it is found, the user can select the concepts corresponding to this assignment and visualise them in GO. If it is not found in GOA, BioMeKe uses HUGO. For a given symbol, HUGO provides synonymous terms and the corresponding SWISSPROT accession number that can be used to access GOA.

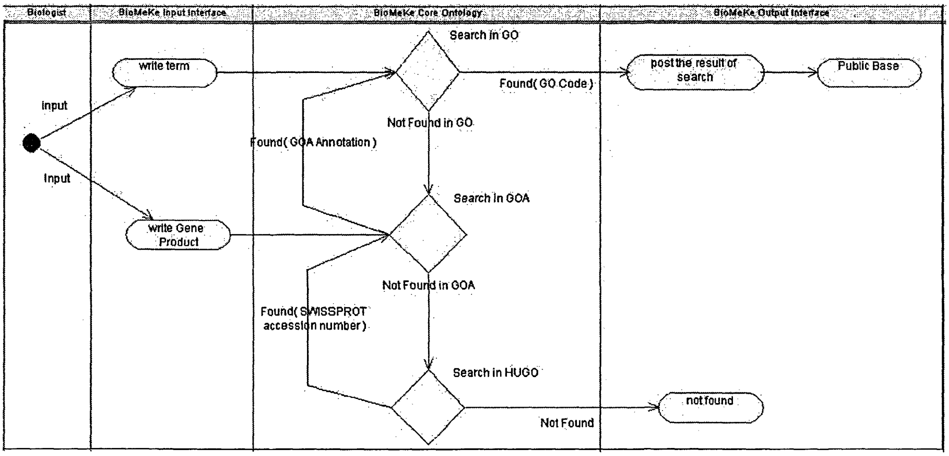
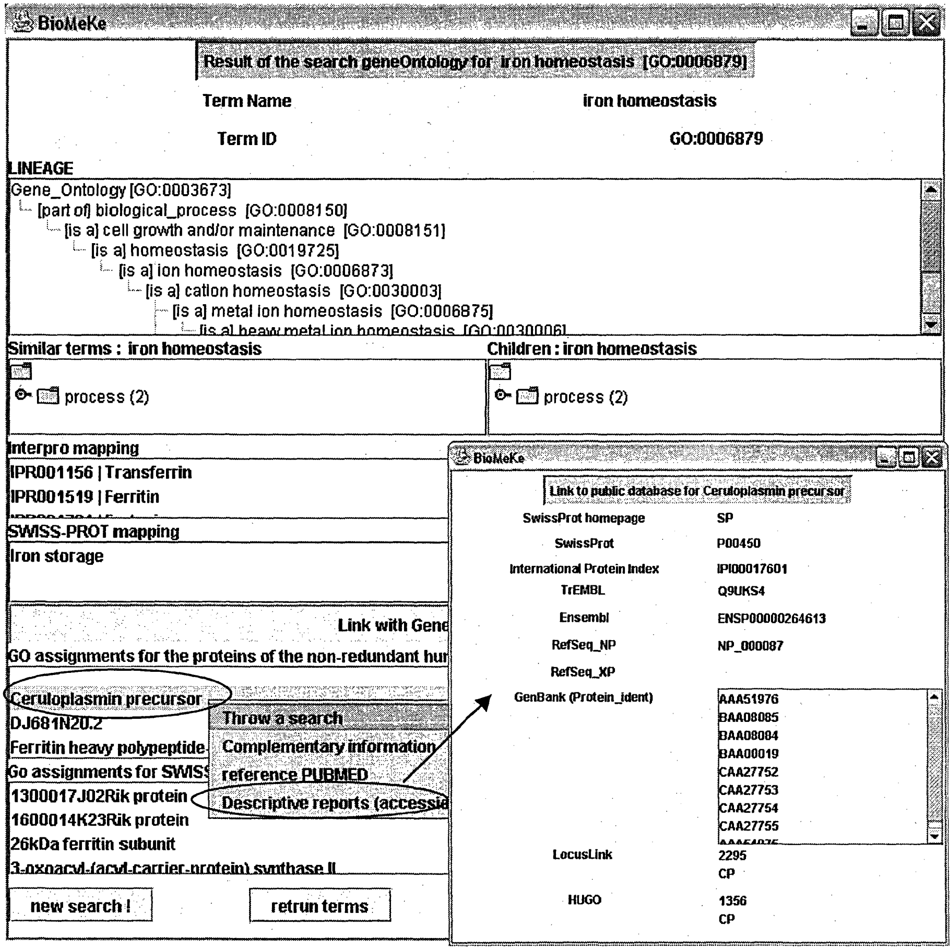


Fig 2 : Search in GO

Link towards public databanks

Many references towards existing public databanks are present in GOA. The ones that are referenced in GOA include SWISSPROT, GenBank, LocusLink, PubMed, InterPro and RefSeq. Therefore, links from GO towards public databanks have been implemented via GOA in BioMeKe. By a click on an accession number in BioMeKe, the appropriate Web page in the public base is opened, providing immediately all the available information (Fig. 3).



BioMeKe

Result of the search geneOntology for iron homeostasis [GO:0006879]

Term Name	iron homeostasis
Term ID	GO:0006879

LINEAGE

```
Gene_Ontology [GO:0003673]
 └ [part of] biological_process [GO:0008150]
    └ [is a] cell growth and/or maintenance [GO:0008151]
       └ [is a] homeostasis [GO:0019725]
          └ [is a] ion homeostasis [GO:0006873]
             └ [is a] cation homeostasis [GO:0030003]
                ├ [is a] metal ion homeostasis [GO:0006875]
                   └ [is a] heavy metal ion homeostasis [GO:0030006]
```

Similar terms : iron homeostasis

process (2)

Children : iron homeostasis

process (2)

Interpro mapping

IPR001156 | Transferrin
IPR001519 | Ferritin

SWISS-PROT mapping

Iron storage

Link with Gene

GO assignments for the proteins of the non-redundant hur

Ceruloplasmin precursor
DJ681N20.2
Ferritin heavy polypeptide

Go assignments for SWISS

1300017J02Rik protein
1600014K23Rik protein
26kDa ferritin subunit
3-oxoacyl-(acyl-carrier-protein) synthase II

BioMeKe

Link to public database for Ceruloplasmin precursor

SwissProt homepage	SP
SwissProt	P00450
International Protein Index	IPI00017601
TrEMBL	Q9UKS4
Ensembl	ENSP00000264613
RefSeq_NP	NP_000087
RefSeq_XP	
GenBank (Protein_ident)	AAA51976 BAA08085 BAA08084 BAA00019 CAA27752 CAA27753 CAA27754 CAA27755
LocusLink	2295 CP
HUGO	1356 CP

Fig 3. Link to public databanks for ceruloplasmin precursor

4. Discussion - Conclusion

BioMeKe allows at this day showing descriptive information on genes and relatives as it is represented in UMLS and GO ontologies, and make possible a unified view and access to a multitude of genomic and medical public databanks by an ontology-based mediation procedure. Data description includes : sequence, chromosome, transcripts, proteins, motifs, and possibly pathologies where they are involved. This ontology-based mediation procedure is set up only for GO.

We choose to use GO ontology, due to its available annotation by the most used biocomputer resources. It is also referenced in other relevant ontologies (MGED ontology and UMLS). A recent study evaluated the UMLS as a terminology and knowledge ressource for bioinformatics by exploring its coverage of terms and relationships needed for molecular biology and genomics [15]. Terms from GeneOntology and LocusLink and all gene products in the GeneOntology Annotation database were mapped to the UMLS. Semantic interoperability will be used for the conceptual annotation . In BioMeKe, it is not set up.

BioMeKe is developed within the context of an ongoing project that seeks to build a gene expression data warehouse for the study of human liver transcriptome in different physio-pathological situations. For each chip's spot among thousands of spots, corresponds a gene or a sequence to which is associated an accession number in GenBank, and probably conceptual annotations in GO and UMLS that need to be extracted systematically[15].

With all this aggregated heterogeneous bioknowledge on liver genes extracted via BioMeKe, and other relevant information that has not been developed in this paper, the next step consists of inferring new biological and medical scenarios using expression data. This includes finding new coregulators and motifs with respect to sequence related data, inferring new gene regulation networks to find out which of the genes are actively regulated and which of them are active regulators, and at mid term certainly pointing the way towards new tools useful for diagnosis or therapeutic actions.

References

[1] Basset DE, Eisen MB, Boguskj MS. Gene expression informatics-it' s all I your mine. *Nat Genet* 1999; 21 (Suppl 1): 51-55.

[2] David DJ, Bittner M, Chen Y, Meltzer P trent JM. Expression profiling using cDNA microarrays. *Nat Genet* 1999; 21 (Suppl 1): 10-14.

[3] Guèrin E, Moussouni F, Courselaud B, Loréal O. UML modeling of Gedaw: A gene expression data warehouse specialised in the liver. *The 3rd french bioinformatics conference proceeding: JOBIM; 2002 June 10-12; France,* Saint-Malo;2002. p. 319-334.

[4] Burgun A, Borenreider O,Le Duff F, Moussouni F, Loréal O. Representation of roles in biomedical ontologies : a case study in functional genomics. Proceedings of *AMIA Annual Symposium* , San Antonio 2002;:86-90.

[5] Marquet G, Burgun A, Moussouni F, Guerin E, Loreal O. An integrative approach of biomedical knowledge via ontologies for liver transcriptome analysis. *Workshop on ontology for biology.* The Studio (Villa Bosch) Heidelberg Novenber 7-8, 2002.

[6] Humphreys BL, Lindberg DA, Schoolman H, Barnette G. The Unified medical Language System : An Informatics Research Collaboration. *J Am Med Inform Assoc.* 1998; 5(1):1-11.

[7] Humphreys BL, Lindgerd DA. The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc* 1993; 81(2): 170-7.

[8] Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. *Pro AMIA Symp* 1999;:181-5

[9] Sperzel WD et al. Biomedical database interconnectivity : an experiment linking MIN,GENBANK, and META-1 via MEDLINE. *Proc Annu Symp Comput Appl Med Care.* 1991; 190-3

[10] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput.* 2000;:517-28.

[11] Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc. 2002 May-Jun;9(3):262-72.

[12] Creating the gene Ontology ressource: design and implementation. *Genome Res* 2001 Aug; 11(8):1425-33

[13] Smith A.D. Oxford Dictionary of Biochemistry and Molecular Biology. *Oxford University Press* 1997

[14] H.M. Wain, M.Lush, F.Ducluzeau, S.Povey. Genew: The Human Gene Nomenclature Database. *Nucleic Acids Research* 2002; Vol.30 : No.1 169-171

[15] Bodenreider O, Mitchell JA, McCray AT: Evaluation of the UMLS as a terminology and knowledge ressource for biomedical informatics. *Proc AMIA Symp* 2002; 61-5

[16] Wu TD. Analysing gene expression data from microarrays to identify candidate genes. *J Pathol* 2001; 195: 53-65.