UMLS-based biomedical annotation of functional genomic data

Gwenaëlle Marquet[1], Emilie Guerin[2], Fouzia Moussouni[2], Olivier Loréal[2] and Anita Burgun[1]

[1] EA 3888, IFR140, Université de Rennes 1 –Faculté de Médecine – 35043 RENNES Cedex - France
{gwenaelle.marquet,anita.burgun}@univ-rennes1.fr
[2] INSERM U522, IFR 140, Université de Rennes 1, CHRU Pontchaillou, 35033 RENNES Cedex -France
{emilie.guerin,fouzia.moussouni,olivier.loreal}@univ-rennes1.fr

Abstract: *The Unified Medical Language System (UMLS) is a potential resource to provide associations between genes and medical knowledge. It may complement GO annotation, which provides information about molecular functions, biological processes, and cellular components associated with genes and gene products. We present the advantages of a UMLS-based annotation (BioMeKE). The annotation method captures the UMLS concepts related to a gene by three types of relations (hierarchical and "other" relations, as well as co-occurrences in Medline) and uses a limited set of 22 Semantic Types for filtering. A set of 43 genes has been used for the evaluation. 100% gene names were mapped successfully to UMLS concepts. The number of concepts that annotated genes in the UMLS was variable, ranging from 0 (for 24 genes) to 673 concepts, after filtering. 63% of the genes had one or more annotations under Disorders and/or Physiology. Among parents and other relations, 80.6% of the information extracted from the UMLS and complementary to GO annotation was expected annotation from the standpoint of our expert. As co-occurrences, most of the information was complementary information, and 40% of the complementary concepts corresponded to expected annotation. The contribution of co-occurrences varies with their frequency in Medline.*

Keywords: annotation, Unified Medical Knowledge System, Gene Ontology, Genew

1 Introduction

Progress in the knowledge of diseases requires a better comprehension of gene-disorders relations. Potential resources include bibliographic databases, e.g. MEDLINE, metabolic pathway databases, e.g. KEGG[1], and specific databases such as Online Mendelian Inheritance in Man™ (OMIM[2]) database [1], which associates gene mutations with the corresponding genetic diseases. Functional annotation is another means to represent the roles that genes and gene products play in biological processes.

The Gene Ontology™ (GO) [2] is a controlled vocabulary for molecular biology and genomics used to annotate gene products in most public databanks. However, GO annotations are limited to Molecular Function, Biological Process, and Cellular Component, and do not provide information on the pathologic conditions that have been associated with genes and their products.

The Unified Medical Language System® (UMLS®) [3, 4] is an "ontology" used in the medical domain that enables physicians to classify signs, symptoms, and diseases using medical concepts. Our hypothesis is that the UMLS may be a powerful tool for providing associations between genes and medical knowledge. Thus, we expect that the combination of GO and UMLS annotation will be a means to relate the modulation of gene expression with expression of diseases in patients, including signs and syndromes.

[1] http://www.genome.jp/kegg/
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Therefore, our goal was to develop a Biological and Medical Knowledge Extractor (BioMeKE) which integrates both ontologies (GO and the UMLS) as well as other public resources, in order to annotate sets of genes with biological and medical concepts. We present the method we have developed. We applied this method to a sample of 43 genes of the iron metabolism. The goal of this work is to evaluate the interest of a medical annotation coupled to a biological annotation and to highlight the bolts at raising on the example of these 43 genes.

2 Resources

The HUGO Gene Nomenclature Committee[3] (HGNC) [5] database, Genew, has been used. For a given gene, different names or identifiers have been proposed by research teams. To manage such heterogeneity, the HGNC proposes nomenclature conventions for genes and now provides approved gene names and symbols, as well as various information, including Swiss-Prot[4] Identifiers (Swiss-Prot ID), LocusLink[5] Identifiers (LocusLink ID).

We used also the November 2004 releases of GO[6] and Gene Ontology Annotation @EBI (GOA[7]). GOA provides assignments of GO terms to gene products for all organisms with completely sequenced genomes, including humans, by a combination of electronic assignment and manual annotation [6].

The UMLS[8] is a general biomedical "ontology" made of two major components, the Metathesaurus®, a large repository of 1,137,344 concepts (2004AC release), and the Semantic Network, a limited network of 135 Semantic Types (Fig. 1).

The Metathesaurus is built by merging more than 100 vocabularies, including Medical Subject Headings (MeSH)[9], Gene Ontology (e.g. GO is a source of "Cell Differentiation" Fig.1) and HUGO. In the Metathesaurus, synonymous terms are clustered under a same concept. Each concept in the Metathesaurus has a Concept Unique Identifier (CUI). For example, the CUI of Transferrin is C0040679. It corresponds to 21 terms. Among them, Transferrin is the preferred term, Transferrins is a variant of the preferred form, Siderophilin and Iron binding protein are synonyms. Within the Metathesaurus, concepts are related by a set of 17,683,827 relations that come from source vocabularies or are acquired during the merging process. The Metathesaurus relationships include Parent (has parent relationship in the Metathesaurus source vocabulary) and Other Relations that can be: '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'[10]. For instance, the concept beta Globulin (C0005157) is a parent of Transferrin, which is related to the concept iron metabolism (C0596803) by an other relation (Fig. 1). In addition, co-occurrences in MEDLINE are stored in the UMLS (Burgun and

[3] http://www.gene.ucl.ac.uk/nomenclature/
[4] http://us.expasy.org/sprot/
[5] http://www.ncbi.nlm.nih.gov/LocusLink
[6] http://www.geneontology.org
[7] http://www.ebi.ac.uk/GOA/
[8] http://www.nlm.nih.gov/research/umls
[9] http://www.nlm.nih.gov/mesh/meshhome.html

Bodenreider 2001). Each co-occurrence has a frequency that is also stored in the UMLS, e.g. Liver neoplasms co-occurs with Transferrin (Fig. 1) with a frequency equal to 3.

The Semantic Network provides a means to categorize all concepts represented in the Metathesaurus. Each Metathesaurus concept is assigned to at least one Semantic Type. For example, the concept Transferrin is assigned to the Semantic Types Amino Acid, Peptide, or Protein and Biologically Active Substance (Fig.1). 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 [7].

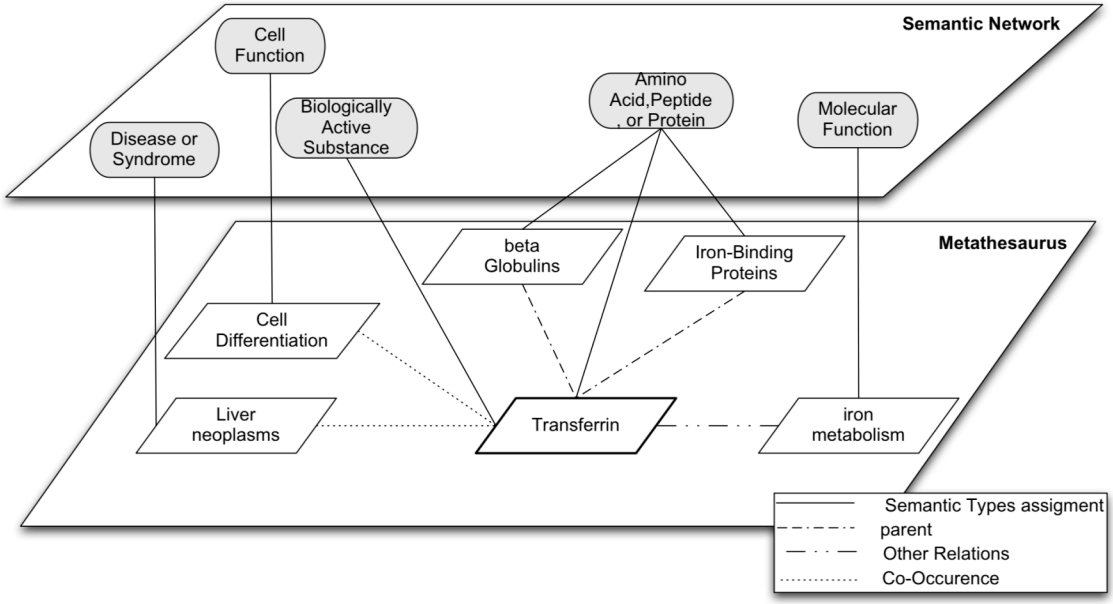

Figure 1. Schematic partial representation of Transferrin concept in the UMLS

3 Methods

5.1 GO annotation

This annotation exploits cross-references in Genew. The Swiss-Prot ID is obtained from the LocusLink ID via Genew. Starting from Swiss-Prot IDs, the list of GO terms that are associated with a gene product are extracted from GOA.

5.2 UMLS annotation

Mapping gene or gene product names to the Metathesaurus: The goal of this step is to extract the initial UMLS concept. Starting from the LocusLink ID, Genew provides the approved name of the gene. It is searched in the Metathesaurus (Fig. 2). Metathesaurus concepts are then filtered using the UMLS Semantic Types: the Metathesaurus concepts assigned to some selected Semantic Types are kept, e.g. the Semantic Type Amino Acid, Peptide, or Protein (Fig.2 Filter 1). The goal of this filtering step is to select only the

[10] http://www.nlm.nih.gov/research/umls/archive/2003AA/UMLSDOC.HTML

Metathesaurus concepts that correspond to genes or gene products. For example, `Ferritin` corresponds to two Metathesaurus concepts, one of them (C0015879) is assigned to the Semantic Type Amino Acid, Peptide, or Protein, the other (C0373607) is assigned to the Semantic Type Laboratory Procedure. The latter is not relevant. The Semantic Type Laboratory Procedure is absent from the list of relevant Semantic Types (Filter 1). Therefore C0015879 is selected whereas C0373607 is not. The output of this step is the set of the Metathesaurus concepts that correspond to the initial LocusLink IDs and are assigned to one of the following Semantic Types: Amino Acid, Peptide, or Protein, Nucleic Acid, Nucleoside, or Nucleotide, Disease or Syndrome, Gene or Genome, Molecular Function (Filter 1). Rationale to keep Disease or Syndrome in that list is that gene names may be similar to disease names, e.g. "Friedreich ataxia 2" (locuslink*:* 2420) or "ATPase, Cu++ transporting, alpha polypeptide (Menkes syndrome)" (locuslink*:* 538)

Searching for Metathesaurus concepts to annotate the gene: After the first process, Metathesaurus relations are explored to perform the annotation. For a given Metathesaurus concept C, the annotation process selects the concepts that are related to C through one of the following relations: PARent (PAR), Other Relations (OR), and Co-OCcurrence (COC) (Related Metathesaurus concepts, in Fig. 2) and are assigned to Semantic Types that may be of interest for the interpretation of postgenomic data in the medical field context. We decided, for this work, to restrict the annotation to only 22 Semantic Types (Filter 2). For example, Cell or Molecular Dysfunction belongs to that list whereas Geographic Area does not. Therefore, given a gene C, a Metathesaurus concept related to C is selected to annotate C if it is assigned to the Semantic Type Cell or Molecular Dysfunction whereas it is not selected if it is assigned to Geographic Area. These 22 Semantic Types are members of seven distinct Semantic Groups: Disorders, Physiology, Chemical & Drug, Genes & Molecular Sequence, Living Being and Phenomena. It is noticeable that some Semantic Types, although potentially interesting "a priori", have not been selected. For example, Body Part, Organ or Organ Component was not chosen by our expert because not really informative due to the fact that it released general information which was not clearly immediately exploitable.

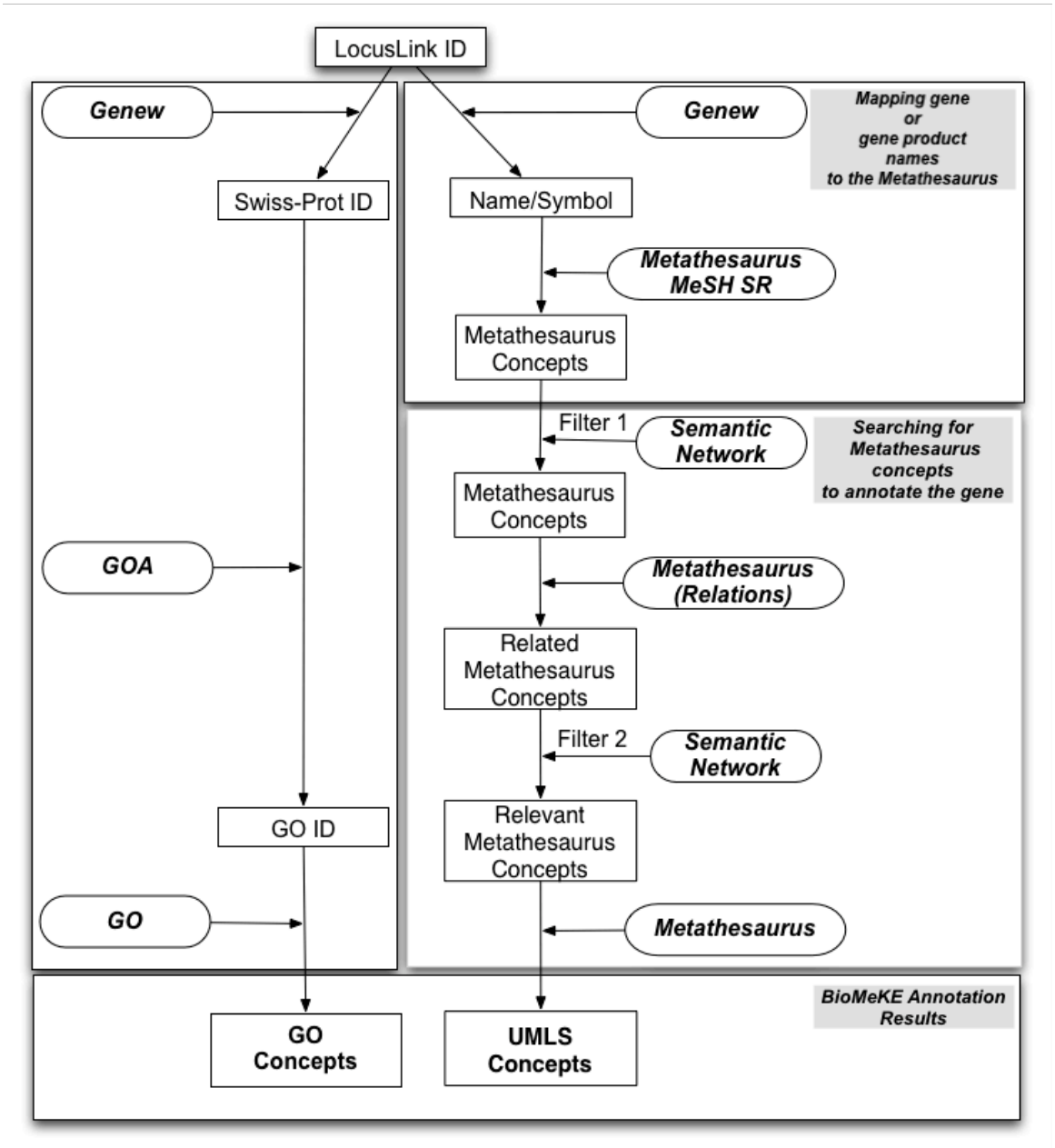

Figure 2. Algorithm of BioMeKE annotation

4 Evaluation

The annotation has been evaluated on a set of 43 genes chosen by U522 biologist experts. Most of these genes are known for being directly or indirectly involved in iron metabolism. Each of the 43 genes has a LocusLink ID that has been recovered via the LocusLink interface.

Several evaluations were implemented:For each gene found in the UMLS, the parents, other related concepts and co-occurring concepts were counted and analyzed per Semantic Type. For co-occurrences, those which were associated with the concept more than ten times were analyzed separately.

1) For each annotated gene, the Related Metathesaurus Concepts restricted to the source GO (UMLS-GO Concepts) have been extracted. To do that, we have considered only the UMLS concepts that are present in GO, namely the Metathesaurus whose sources include GO terminology. The redundancy between the UMLS-GO Concepts and the GO annotation has been evaluated.

2) Evaluation of the biomedical interest by an expert of the domain (OL). Two criteria were used. The first one was whether the information was considered complementary to GO annotation (complementary annotation) or not (not complementary annotation). An UMLS annotation is regarded as not complementary annotation compared to GO when our expert considers that an UMLS annotation and a GO annotation have the same meaning, e.g. for the gene "lactotransferrin", among the GO annotations, there is the GO term 'ferric iron binding' while in the UMLS annotation we found the concept 'Iron-binding Proteins'. This concept is not a complementary annotation for our expert; it is a not complementary annotation.

The second criterion was the relevance of the concept, with two subclasses (expected and not expected). This criterion was evaluated only on the annotation that is judged complementary to the first criteria:

- Expected information, which was expected from the standpoint of our expert e.g. Iron Overload is expected complementary information for Transferrin annotation
- Not expected information, which was information previously unknown from the medical and biological knowledge of the expert e.g. Vitamin A deficiency was qualified not expected for transferrin annotation.

5 Results

Among the 43 genes that we used for the evaluation, 100% were found in the UMLS. Twenty eight genes of the 43 were mapped to a single UMLS concept, 13 were mapped to two concepts e.g. "lactotransferrin" (LocusLink: 4057) mapped with C0022942 and C1416933, and two were mapped to three concept.

Most of the gene concepts found in the UMLS (42) were assigned in the UMLS to the Semantic Type Gene or Genome. Two genes were also assigned in more to Disease or Syndrome e.g. "hemochromatosis": LocusLink 3077 and "Von Hippel-lindau tumor supessor": LocusLink 7428.

Among the 43 genes that were present in the UMLS, 19 genes had annotation, i.e. concepts related to them by parent, other relations or co-occurrence were found. The number of concepts that annotated genes in the UMLS was variable, ranging from 1 to 673 concepts, after the filtering process based on the 22 Semantic Types that were previously qualified relevant. The type of relations between the annotating concepts to the annotated gene influences the number of concepts release (Fig. 3). The number of annotating concepts is maximal for co-occurrences with a frequency equal to or less than 10 in MEDLINE (C1 relation in Fig. 3).

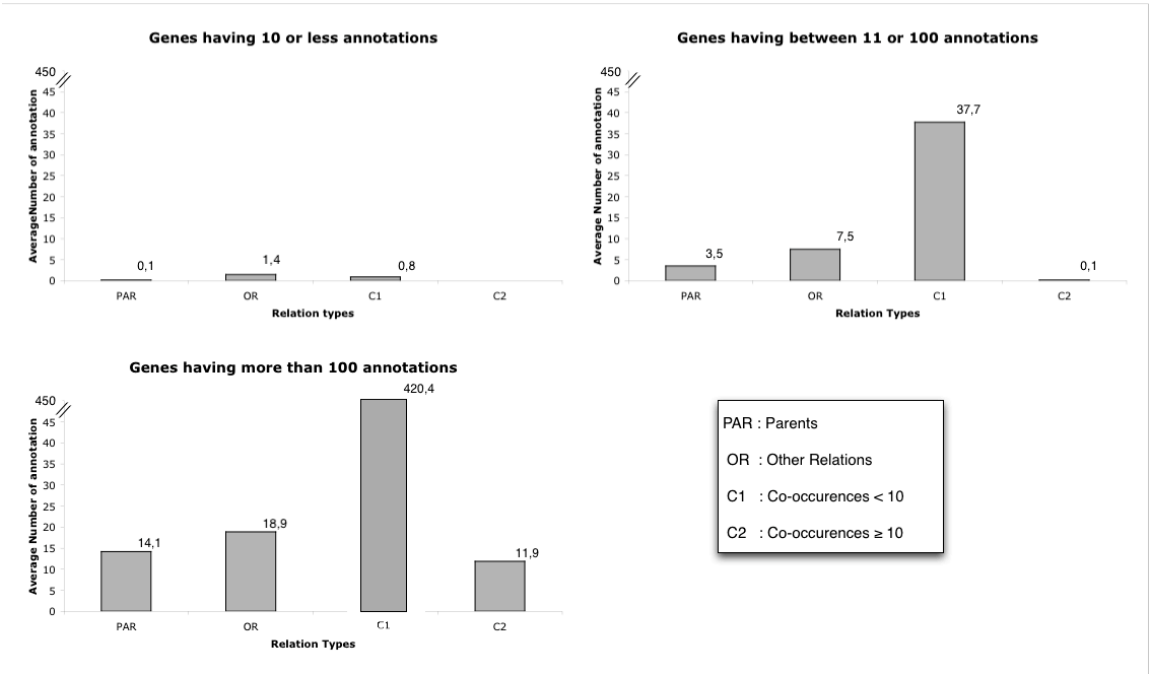

Figure 3. Representation of the average number of annotations per relation types.

In order to know if the annotations obtained from the UMLS give clinical or phenotypic information, we analysed the frequency of annotations per Semantic Group. Figure 4 presents the repartition of the annotating concepts, after filtering.

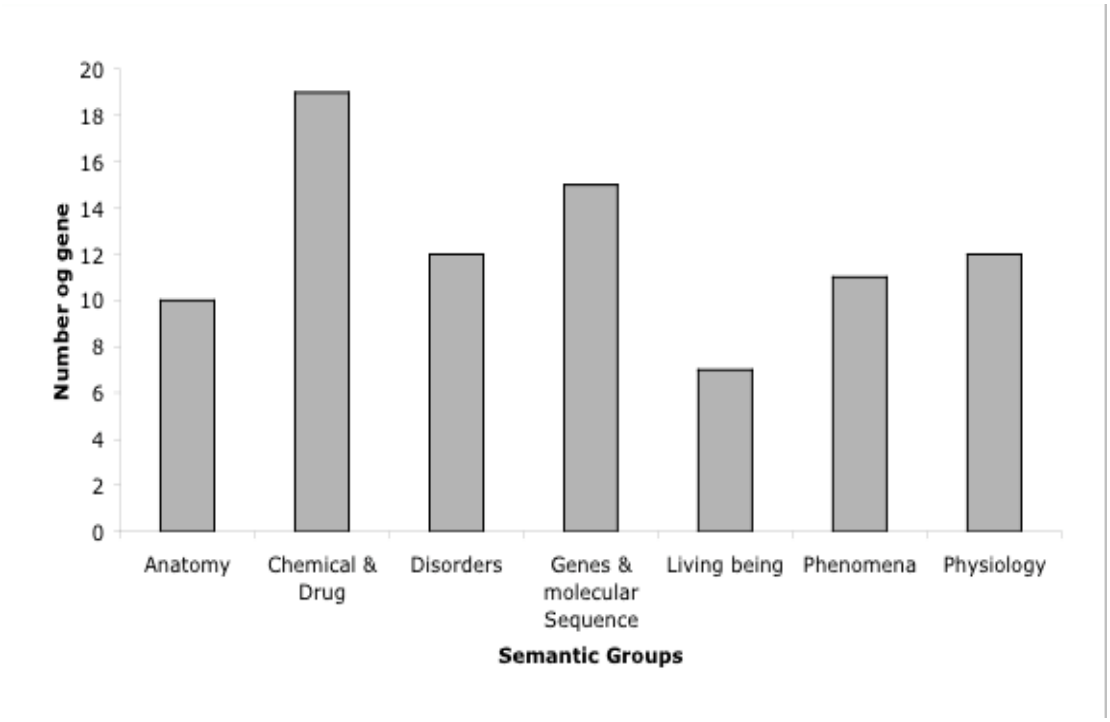

Figure 4. Number of genes which have at least one annotation under the 7 Semantic Groups.

All the 19 genes annotated by the UMLS had one or more annotations under Chemicals & Drugs, mostly represented in our sample by the Semantic Type Amino Acid, Peptide or Protein. 63% of the genes had one or more annotations under Disorders and/or Physiology. 94% of the concepts which annotated the 19 genes are categorized under the three following Semantic Groups: Chemicals & Drug, Disorders and Physiology. The UMLS annotation of Transferrin has been chosen as an example and is illustrated in Table 1. Iron binding Proteins, Beta Globulin, Metalloproteins, Transferrin Receptor, and Hemochromatosis represent parents, other relations, and co-occurrences at high frequency and were expected by the expert as annotating concepts for Transferrin. Among the co-occurring concepts with low frequency, the only one that was expected was beta-Thalessemia due to the fact that this disease induces the development of a secondary hemochromatosis (Table 4).

UMLS Concepts	**Relation Types**
Chemicals & Drugs	
Iron-Binding Proteins	Parent
beta Globulin	Parent
Metalloproteins	Other Relations
Transferrin Receptor	Other Relations
Carrier Proteins	Co-occurrence (freq: 25)
Ceruloplasmin	Co-occurrence (freq: 11)
Chorionic Gonadotropin	Co-occurrence (freq: 1)
Disorders	
Hemochromatosis	Co-occurrence (freq: 23)
beta-Thalassemia	Co-occurrence (freq:1)
Arteriosclerosis	Co-occurrence (freq: 1)
Lung diseases	Co-occurrence (freq: 1)
Hemochromatosis	Co-occurrence (freq: 23)

Table 1: A part of UMLS annotation for Transferrin (LocusLink ID: 7018).

5.1 Comparison UMLS Annotation of medical annotation (UMLS annotation) and biological annotation (GO annotation)

Of all the 4547 Related Metathesaurus Concepts, 560 (12.3%) are UMLS-GO concepts. Redundant concepts between the medical annotation and biological annotation represent 0.5% of the UMLS-GO Concepts and 2% of the GO Concepts.

Among the 19 annotated genes, two genes had one or more annotations that were provided by both medical annotation (UMLS-GO Concept) and biological annotation (GO Concept). These two genes are:

- The gene "glyceraldehyde-3-phosphate dehydrogenase" (LocusLink: 2597), which is annotated by 23 UMLS-GO Concepts and by five GO Concepts. It has one shared annotation: glycolysis
- The gene "erythropoeitin receptor" (LocusLink: 2057), which is annotated by 47 UMLS-GO Concepts and by three GO Concept. It has two shared annotations: erythropoietin receptor activity and signal transduction

These three annotations are categorized under the Semantic Type Molecular Function.

5.2 Contribution of the medical annotation

For each gene, the annotations per relation were counted and the expert evaluated the complementary information compared to GO (Figure 5).

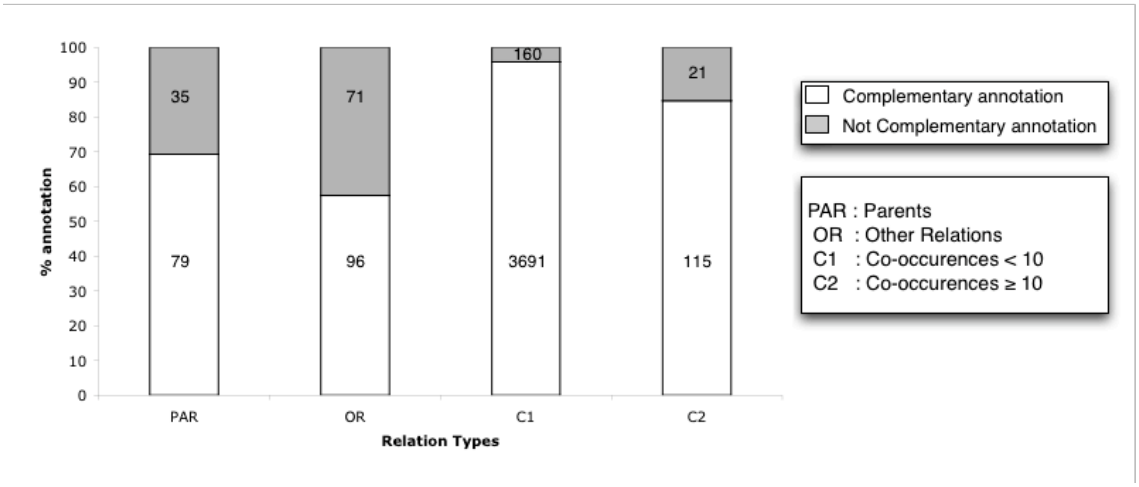

Figure 5. Repartition of annotating concepts per category (Complementary in GO annotation or not complementary in GO annotation) for each relation type. The number of concepts is given in the box.

Among parents and other relations, 63.4% of the concepts correspond to complementary annotation.

For co-occurrences, we analyzed the co-occurrences with frequency < 10 on the one hand and co-occurrences with frequency ≥10 on the other hand. For the co-occurrences with a frequency < 10, we found that only 4.2% of the annotating concepts were not complementary annotation to GO annotation. For co-occurrences with frequency ≥ 10, 86.7% of the annotating concepts were complementary to GO annotation.

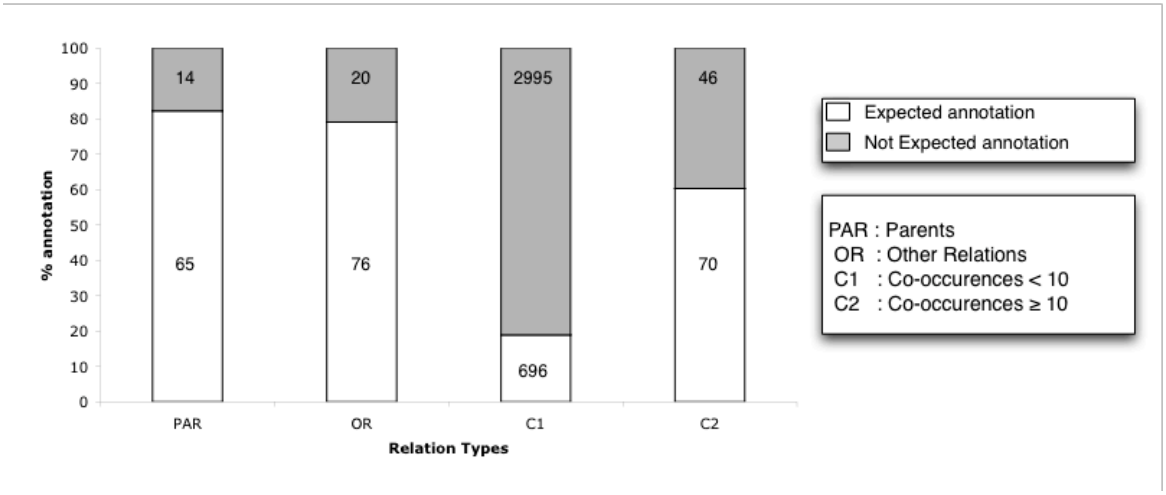

Figure 6. Repartition of complementary annotation concepts per category, (expected / not expected) for each relation type. The number of concepts is given in the box.

Among parents and other relations, 80.6% of complementary annotation were judged expected annotation (141 out of 175 concepts).

For co-occurrences with frequency < 10, 18.8% of the complementary annotation corresponded to expected complementary annotation and 81.1% corresponded to not expected complementary annotation. For co-occurrences with frequency ≥ 10, 60% corresponded to expected complementary annotation, and 40% to annotation that was not expected by the expert.

6 Discussion

BioMeKE has been built in order to have an integrated system for a biological and medical annotation. We presented the methods that we used and its evaluation on a set of 43 genes.

The originality of BioMeKE is to perform both medical and biological annotation. Every genes in our test set are annotated by both biological concepts (GO concepts) and medical concepts (UMLS concepts).

The UMLS as it is used in BioMeKE provides concepts that are related to the gene in existing terminologies. Genew was integrated in the UMLS. We used this integration for relating genes to corresponding gene concepts in the UMLS.

The UMLS consists of two major components: the Metathesaurus and the Semantic Network. BioMeKE used these components to product a medical annotation. In fact, the Metathesaurus provides related concepts to genes by different relations: Parents, Other relation and Co-occurrences in MEDLINE. The Semantic Network is used for filtering tasks [9]. The first filter selects only the Metathesaurus concepts that correspond to genes or gene products, and the second filter provides annotation which may be of interest for the interpretation of postgenomic data in the medical field context.

The Metathesaurus merges 100 vocabularies including a medical resource containing genetic diseases related to gene mutation: OMIM. In fact, some annotating UMLS concepts correspond to OMIM

terms, e.g. annotating UMLS concepts for Transferrin, is Insulin-Like Growth Factor II, whose sources include OMIM. The Metathesaurus include also Gene Ontology [10]. However, it is important to provide two different annotation processes for GO annotation and for UMLS annotation. Indeed, UMLS annotation includes GO terms associated with genes in the UMLS Metathesaurus; these two annotations (UMLS and GO annotation) have been different. For example, BioMeKE gives 84 UMLS-GO Concepts (§4) (Endocytose, cell Differentiation, etc) but all are different from GO annotations (Ferric iron binding, Iron ion homeostasis, Iron ion transport, Transport, Extracellular region). We have shown that the overlap between the medical annotation and the biological annotation is very limited. Indeed, on the set of genes, only two genes shared one or two annotations between medical annotation and biological annotation.

Like BioMeKE, GenesTaces uses the UMLS [11]. GenesTrace map UMLS concepts (disease) to genes via GO and LocusLink. This tool appears to be complementary to BioMeKE. Starting from a gene G, for an association between G and a disease D provided by our annotation, GenesTrace searches for other genes that are related to D.

The heterogeneity of gene identifiers, as already evoked by some authors [12, 13, 14] is a bottleneck of biomedical annotation. Several resources are used by BioMeKE to overcome heterogeneity in gene naming. Genew provides explicitly the relationship between genes and proteins as well as several cross-references. Genew gives also the official name of genes and it is used by BioMeKE to manage the heterogeneity of names and gene identifiers. However, we have noted that an official name can be associated to more than one UMLS concept, for example the "ATPase, Cu++ transporting, beta polypeptide (Wilson disease)" gene (Symbol: ATP7A, LocusLink: 540) is mapped to the UMLS concept C0296649 and the UMLS concept C1412689.

These two concepts correspond to the same string "ATPase, Cu++ transporting, beta polypeptide (Wilson disease)" but their sources and their synonyms are different. MeSH is the source of the first concept and HUGO is the source of the second one. Among their synonyms, the term "Wilson disease" is a synonym of the first concept and "ATP7A protein" is a synonym of the second. One another barrier is polysemy. A filtering process (Filter 1, fig 2) is required to select only the terms that correspond to genes. The filtering relies on the UMLS Semantic Types, which are used to categorize all Metathesaurus concepts, e.g. `Transferrin` corresponds to three Metathesaurus concepts, one of them (C0040679) is assigned to the Semantic Type Amino Acid, Peptide, or Protein, another C1442762 is assigned to the Semantic Type Gene or Genome and C0202105 is assigned to the Semantic type Laboratory Procedure. The latter is not relevant.

BioMeKE uses official Genew name. Therefore, it is possible that more information could be obtained using other names. For example, BioMeKE exploits the official name present in Genew but does not take advantage of previous gene names. In future works, we will to implement more paths in BioMeKE.

Despite this limitation, the UMLS annotation was considered interesting by our expert. Indeed, from the UMLS evaluation, 76.8% of the UMLS annotation gives complementary annotation compared to GO and 60.1 % of this complementary information was judged relevant by the expert. For the Parent and Other relations, 80.6% of complementary annotation are expected and for the co-occurrences ≥ 10, 60% of

complementary annotation are also expected. The parents, other relations, and co-occurrences with frequency ≥ 10 give valid information for non experts of the domain and every the co-occurrences can be used for suggesting new hypotheses for biologists and physicians.

Acknowledgements

This project is supported by Région Bretagne (20046805 , PRIR 139).

References

[1] Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res 30*(1):52-5.

[2] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT and others. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25-9.

[3] Bodenreider O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res 32 Database issue*:D267-70.

[4] Lindberg DA, Humphreys BL, McCray AT. 1993. The Unified Medical Language System. *Methods Inf Med 32*(4):281-91.

[5] Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S. 2004. Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res 32 Database issue*:D255-7.

[6] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res 32 Database issue*:D262-6.

[7] Burgun A, Bodenreider O. 2001. Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. *Medinfo. 2001*;10(Pt 1):171-5.

[8] McCray AT, Burgun A, Bodenreider O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo. 2001*;10(Pt 1):216-20

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

[10] Lomax J, McCray AT. 2004. Mapping the Gene Ontology into the Unified Medical Language System. *Comparative and Functional Genomics* 5(4):354-361.

[11] Cantor MN, sarkar IN, Bodenreider O, Lussier YA. GenesTrace*:* Phenomic knowledge discovery via strutured terminology. *Pacific Symposium on Biocomputing 2005* p103-114

[12] Karp, P.D. Database Links are a Foundation for Interoperability. *Trends in Biotechnology*, vol. 14, pp. 273-279, 1996

[13] Etzold T, Ulyanov A, Argos P. SRS: Information Retrievial System for Molecular Biology Data Banks. *Methods in Enzymology*, 266:114-28,1996.

[14] Stevens R, Goble C, Paton N, Bechhofer S, Ng G, Baker P, Brass A: Complex Query Formulation Over Diverse Information Sources Using an Ontology. In Zoe Lacroix and Terence Critchlow, editors, *Bioinformatics: Managing Scientific Data*. Morgan Kaufmann, May 2003.