Mapping protein information to disease terminologies

Anaïs Mottaz[1], Yum L. Yip[1,2], Patrick Ruch[2,3], and Anne-Lise Veuthey[1]

[1] Swiss Institute of Bioinformatics,
[2] Dept. of Structural Biology and Bioinformatics, University of Geneva,
[3] Medical Informatics Service, University Hospitals of Geneva,
Geneva, Switzerland

Abstract

In order to improve the accessibility of genomic and proteomic information to medical researchers, we have developed a procedure to link biological information on proteins involved in diseases to the MeSH and ICD-10 disease terminologies. For this purpose, we took advantage of the manually curated disease annotations in more than 2,000 human protein entries of the UniProt KnowledgeBase. We mapped disease names extracted from the entry comment lines or from the corresponding OMIM entry to the MeSH. The method was assessed on a benchmark set of 200 manually mapped disease comment lines. We obtained a recall of 54% for 91% precision. The same procedure was used to map the more than 3,000 diseases in Swiss-Prot to MeSH with comparable efficiency. Tested on ICD-10, the coverage of the mapped terms was lower, which could be explained by the coarse-grained structure of this terminology for hereditary disease description. The mapping is provided as supplementary material at http://research.isb-sib.ch/unimed.

1 Introduction

With the emergence of high-throughput technologies, the amount of biomedical data available to researchers and clinicians has increased drastically over the last decade. In the genomic/proteomic era, new methods of knowledge management will soon allow researchers to move beyond the analysis of single molecule or pathway to consider global mechanisms, such as pathological processes, from an integrated point of view. One of the challenges for bioinfomatics in this context is to bridge the gap between biological knowledge and clinical data. Currently, the main obstacle to achieve this objective is the compartmentalization of the data in different databases, and the inconsistencies in the vocabulary used by these resources to describe biomedical entities and concepts. A key solution to this interoperability problem lies in the development of common terminologies capable of acting as a metadata layer to provide the missing links between the various resources. Successful initiatives for the development of standardized vocabularies in the biological domain started some years ago with the creation of the Gene Ontology (GO) for the description of biological functions and processes [1]. It was followed by the developments of numerous biological ontologies under the Open Biological Ontologies initiative (OBO) [2]. In the medical domain, the effort on the development of standard terminologies started many years before these initiatives in the molecular biology domain. Key vocabularies such as the International Classification of Diseases (ICD) [3], the SNOMED clinical terminology [4], and the Medical Subject Headings (MeSH) [5] were developed in order to standardize information on various domains of medicine, from patient care to biomedical literature indexing. The Unified Medical Language System (UMLS) [6] was developed by the US National Library of Medicine (NLM) to function as an umbrella over these resources by providing a system of interrelations between all these terminologies.

Even if the recent integration of GO in the UMLS has opened new ways of linking biological and medical resources via terminologies, relationships between gene functions and diseases are still poorly documented in terminologies. Several initiatives have been set up to link phenotypes to genotypes [7], and systems have been developed to detect such associations. For instance, GenesTrace™ [8] and BioMeKe [9] use the relationships between GO and UMLS concepts of disease-related semantic types to infer gene-disease relationships. PhenoGO uses natural language processing methods to assign phenotypic context to GO annotations [10]. The MedGene database gathers relationships between human gene names and diseases extracted from MEDLINE [11]. GFINDer uses textual information from the Online Meendelian Inheritance in Man database (OMIM) to analyze correlation of disease with gene expression in microarray results [12]. All these systems rely on inference and, therefore, depend closely upon the accuracy of the various methods. A straighter way to link genes to diseases would be to use the disease-related information directly provided by some specific biological databases. Take the example of the UniProt Knowledgebase (UniProtKB) [13], the most comprehensive protein warehouse with extensive annotation and cross-references to other database resources. In UniProtKB, more than 2,000 human proteins contain manually curated information related to their involvement in pathologies. This information comes with the type and position of the single amino acid polymorphisms known to cause the disease, and cross-references to variant databases and genomic resources, such as dbSNP and Ensembl. While this information is clearly of value, it is not easily accessible for clinical researchers due to the fact that UniProtKB does not use standard medical vocabularies to describe diseases associated to proteins and their variants.

In this study, we have developed an automatic approach to map the disease terms in UniProtKB to two well-known and widely used disease terminologies within the UMLS: MeSH - the controlled vocabulary thesaurus used for biomedical and health-related documents indexing [5], and ICD-10 - the official disease classification provided by the WHO [3]. We took advantage of the manual annotation in UniProtKB as well as the curated links of UniProtKB entries to OMIM, the comprehensive knowledge base of human genes and genetic diseases [14]. A benchmark set was created for the refinement of term matching algorithm as well as for the definition of matching score and score threshold. This work provides a basis for further work aiming to increase the interoperability between data resources from the medical informatics and the bioinformatics domains.

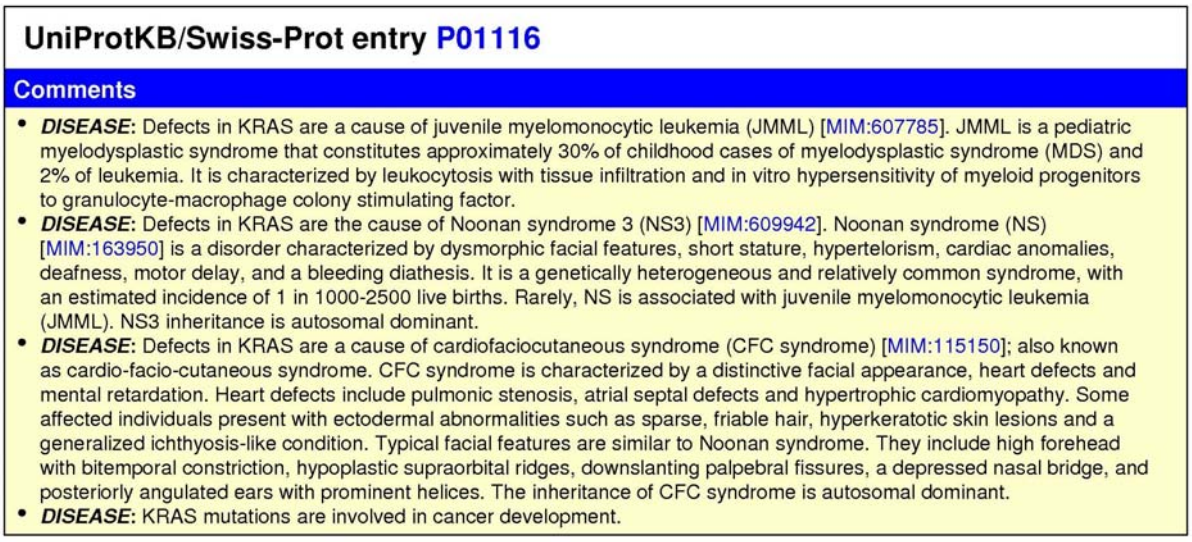
2 Methods

2.1 Extraction of disease names

The UniProtKB/Swiss-Prot (release 52.5), and the OMIM (version May 2007) were used for this study. In UniProtKB/Swiss-Prot entries, disease information related to the protein is expressed in free text comment lines qualified by the category ‘Disease’ (Figure 1). By manual inspection, we first established a list of regular expressions that indicates the presence of a disease name within these lines (e.g. ‘cause(s)’, ‘cause of’, ‘involved in’, ‘contribute(s) to’, ‘induce(s)’). The disease name was usually delimited either by the end of a sentence, a conjunction or relative clause, or by the corresponding OMIM identifier. We also defined a list of specific words, such as ‘susceptibility to’, ‘development (of)’, ‘various types of’ to remove terms that have no direct connection with the disease name. In rare cases where several diseases were described in the same comment line, we restricted the extraction to the first mentioned disease.

In parallel, we took advantage of the citations to OMIM phenotypes (#) and genes with phenotypes (+) in the disease comment lines to extract the fields *Title* and *Alternative titles; symbols* from the corresponding OMIM entries. These two fields provide the disease name in

OMIM as well as a set of synonyms. For names coming from "gene with phenotype (+)" entries, we did not try to distinguish between gene names and diseases names, both types were included in the disease list.



UniProtKB/Swiss-Prot entry P01116

Comments

- ***DISEASE***: Defects in KRAS are a cause of juvenile myelomonocytic leukemia (JMML) [MIM:607785]. JMML is a pediatric myelodysplastic syndrome that constitutes approximately 30% of childhood cases of myelodysplastic syndrome (MDS) and 2% of leukemia. It is characterized by leukocytosis with tissue infiltration and in vitro hypersensitivity of myeloid progenitors to granulocyte-macrophage colony stimulating factor.
- ***DISEASE***: Defects in KRAS are the cause of Noonan syndrome 3 (NS3) [MIM:609942]. Noonan syndrome (NS) [MIM:163950] is a disorder characterized by dysmorphic facial features, short stature, hypertelorism, cardiac anomalies, deafness, motor delay, and a bleeding diathesis. It is a genetically heterogeneous and relatively common syndrome, with an estimated incidence of 1 in 1000-2500 live births. Rarely, NS is associated with juvenile myelomonocytic leukemia (JMML). NS3 inheritance is autosomal dominant.
- ***DISEASE***: Defects in KRAS are a cause of cardiofaciocutaneous syndrome (CFC syndrome) [MIM:115150]; also known as cardio-facio-cutaneous syndrome. CFC syndrome is characterized by a distinctive facial appearance, heart defects and mental retardation. Heart defects include pulmonic stenosis, atrial septal defects and hypertrophic cardiomyopathy. Some affected individuals present with ectodermal abnormalities such as sparse, friable hair, hyperkeratotic skin lesions and a generalized ichthyosis-like condition. Typical facial features are similar to Noonan syndrome. They include high forehead with bitemporal constriction, hypoplastic supraorbital ridges, downslanting palpebral fissures, a depressed nasal bridge, and posteriorly angulated ears with prominent helices. The inheritance of CFC syndrome is autosomal dominant.
- ***DISEASE***: KRAS mutations are involved in cancer development.

Figure 1: disease comment lines in a UniProtKB/Swiss-Prot entry

2.2 Mapping procedure

We mapped the extracted disease names to the terms from the disease category of the MeSH terminology (version 2007). The MeSH thesaurus is structured in a hierarchy of descriptors, each descriptor including a set of related concepts, and each concept itself containing a set of terms, which are synonyms and lexical variants. We mapped the disease names to the MeSH terms and linked the results to the corresponding MeSH descriptors. For ICD-10, we mapped the disease names to all non-redundant terms of ICD-10, without distinction of their types.

The mapping procedure consisted of two successive term matching steps:

(1) we found exact matches, where all words composing the name had an identical correspondent in a MeSH term and vice versa, the word order and the case not being taken into consideration.

(2) in case of no exact match, we looked for partial matches by decomposing the name into its word components and calculated a similarity score for names having at least one word in common.

The score used to determine the similarity between two terms was calculated as a function of the number of words in common minus the number of words that differ. In order to take into account the informative content of each word composing the term, we weighted them according to an adaptation of the weighting schema 'Term Frequency x Inverse Document Frequency' (TF X IDF), commonly used in information retrieval techniques [15]. We calculated the inverse document frequency (IDF) of each word present in the three sources of terms, namely Swiss-Prot disease lines, OMIM *Titles* and *Alternative titles*, and disease MeSH terms or ICD-10 terms. The similarity score was calculated according to the following formula:

$$S = \frac{\sum_{cw} \log_2\left(\frac{1}{freq(cw)}\right) - \sum_{ncw} \log_2\left(\frac{1}{freq(ncw)}\right)}{size(disease)}$$

Where *freq=n/N*, with *n* the number of occurrence of the word in all OMIM (Titles, Alternative titles), Swiss-Prot disease comment lines, and MeSH terms (disease category) or ICD-10 terms. *N* represents the total number of words in these documents. *cw* stands for words in common and *ncw* for words present in only one of the terms. The term *size(disease)* is a normalization factor consisting of the number of words composing the disease name to be mapped.
Hyphenated words were treated in a special way to avoid false positive matches without penalizing the sensitivity. Each of their components was considered as distinct word. If all components had a matched equivalent, their respective weights were summed up in the score calculation. Otherwise, their weights were subtracted.

2.3 Mapping evaluation

In order to evaluate the mapping procedure, 200 disease comments from 97 UniProtKB/Swiss-Prot entries were manually mapped to MeSH by a medical expert. Swiss-Prot entries were selected randomly. However, care was taken so that the chosen sample of entries would be representative and lead to a proportion of exact and partial matches similar to that found in a preliminary mapping attempt. The disease terms were mapped, whenever possible, to a single MeSH term of the same granularity or close in the hierarchy. However, when no equivalent term was found in the terminology, the disease name was mapped to several parents in different hierarchies or to high level concepts.
The mapping procedure was assessed in terms of *precision=TP/(TP+FP)* and *recall=TP/total number of terms*, where *TP* is the number of correct mappings (true positives), and *FP* the number of incorrect mappings (false positives).

3 Results

In UniProtKB/Swiss-Prot (release 52.5), 2,167 human protein entries contained information on the involvement of these proteins in diseases. This corresponded to a total of 3,197 diseases, mainly of genetic causes. Among these diseases, 2,410 had a link to a corresponding phenotype described in OMIM, which represented 77% of the total OMIM entries of phenotypes with a known molecular basis (version May. 2007). We mapped the disease names to the 38,646 terms of the MeSH disease category (version 2007) and 29,550 non-redundant terms of ICD-10. We treated independently names provided by Swiss-Prot and those provided by OMIM. A benchmark set consisting of 200 disease comment lines with 173 references to OMIM was used to evaluate the mapping procedure.

3.1 Disease name extraction

Swiss-Prot disease names were extracted from the comment lines with a set of regular expressions. As the Swiss-Prot disease lines are usually well structured, we were able to extract almost all disease names. The extraction failed in only 7 comment lines where a clear reference to a disease was not expressed, for instance:

"(CBL) can be converted to an oncogenic protein by deletions or mutations that disturb its ability to down-regulate RTKs." (P22681)

The system was constructed to extract only a single disease name per line. By manual assessment of the extraction results, we noticed that in some cases it failed to treat correctly lines such as:

"KRT16 and KRT17 are coexpressed only in pathological situations such as metaplasias and carcinomas of the uterine cervix and in psoriasis vulgaris." (P08779)

We did not investigate further these cases, as the structure of disease lines is planned for a revision in the framework of Swiss-Prot comment standardization efforts.

Extraction of OMIM's disease names from *Title* and *Alternative title; symbols* was simple. We kept all words composing a term, except qualifiers such as "included" or "obsolete".

3.2 Mapping on the benchmark

The results from a benchmark of 200 diseases manually mapped to MeSH terms are shown in Table 1. The mapping was done independently on disease names extracted from Swiss-Prot and on *Title* or *Alternative titles* of OMIM.

The mapping procedure was divided into two successive steps. First, we checked for exact matches with MeSH terms. Exact matches covered about 20% of the benchmark with an excellent precision. The only three false positive matches were caused by a difference of classification between MeSH and OMIM. More specifically, OMIM considers these terms as synonyms, whereas MeSH classified them in different concepts. For instance, two types of *epidermolysis bullosa*, which are distinct MeSH descriptors, are synonyms in OMIM. When we gathered the exact matches provided by the two resources, the coverage increased to 26%.

The terms without exact matches went through a partial matching procedure where the similarity between terms was measured according to a score derived from information retrieval scoring techniques. The score threshold for taking the mapping into consideration was set at 0. This appeared to be a good trade-off between recall and precision (Figure 2), given that precision is an essential requirement for an automatic mapping procedure. Therefore, the recall is rather low (24%) but the precision is good (86-89%). Adding the partial mappings of Swiss-Prot and OMIM increased the recall to 29% with the same precision.

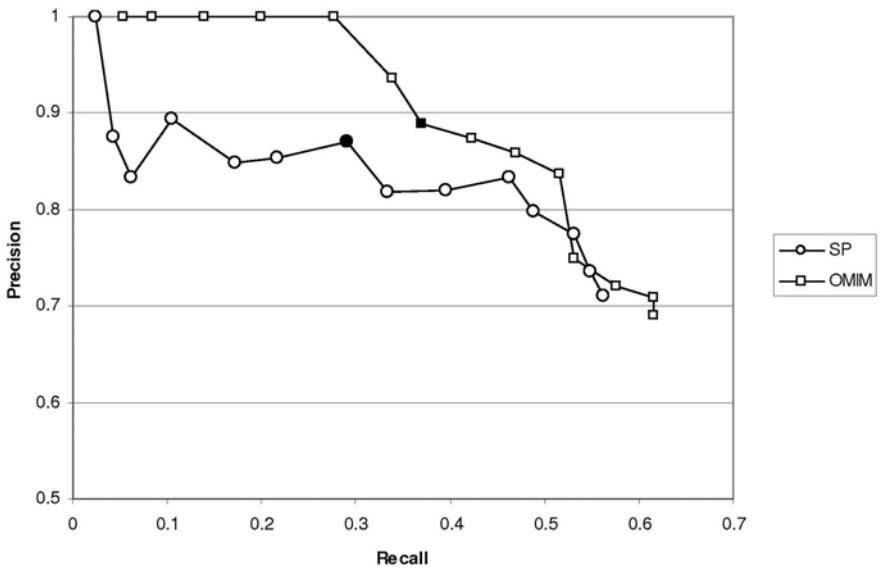


Figure 2: Recall-precision at each integer unit of the similarity score in the interval [-7,+7]. The black points correspond to the recall/precision at the selected score threshold (0).

The final performance of the system was measured by adding exact and partial matches from the two resources. In this case, we obtained a recall of 55% for a precision of 92%. In order to increase the confidence of the mapping, we also calculated the performance when we considered the union of exact matches and the intersection of partial matches. For the latter, we counted only matches where the two sources (SP and OMIM) pointed to the same MeSH

descriptor. In this case, the number of partial matches from both sources was half reduced, but the precision increased to 95%, for a global recall of 39%.

Table 1: Result of the mapping of 200 disease comment lines from Swiss-Prot on MeSH terms. SP∩OMIM means that both mappings correspond to the same MeSH descriptor.

	Exact match			Partial match			Total		
	Retrieval	Recall	Precision	Retrieval	Recall	Precision	Retrieval	Recall	Precision
SP	35 (18%)	35 (18%)	100%	54 (27%)	47 (24%)	87%	89 (45%)	82 (41%)	92%
OMIM	43 (22%)	40 (20%)	93%	54 (27%)	48 (24%)	89%	97 (49%)	88 (44%)	91%
SP ∩ OMIM	23 (12%)	23 (12%)	100%	28 (14%)	26 (13%)	93%	62 (31%)	60 (30%)	97%
SP ∪ OMIM	54 (27%)	52 (26%)	96%	64 (32%)	57 (29%)	89%	118 (59%)	109 (55%)	92%

SP: Swiss-Prot

3.3 MeSH and ICD-10 mapping to UniProtKB/Swiss-Prot disease comment lines

We applied the mapping procedure set up with the benchmark to a total number of 3197 disease comment lines present in Swiss-Prot, with 75 % of them having a corresponding OMIM entry. The mapping was performed on both MeSH and ICD-10 terminologies and the results are detailed in table 2. We extrapolated the same procedure to map ICD-10, even if a condition such as the score threshold was not assessed on this terminology.

In term of retrieval, i.e. the number of disease names matching a term above the threshold, the results of the mapping with MeSH were just slightly lower compared with those of the benchmark. The lower coverage of the OMIM mapping could be explained by the fact that the proportion of Swiss-Prot diseases with OMIM cross-references was higher in the benchmark (86% instead of 75%). Considering matches with Swiss-Prot and OMIM terms, 54% of the disease comment lines were mapped.

It was not possible to measure the performance of the system in terms of precision and recall. As a first assessment of the mapping, we simply checked if, in case of exact matches, corresponding Swiss-Prot and OMIM terms mapped to identical MeSH descriptors. This was confirmed in all but 12 cases. One case was due to a problem of multiple diseases mentioned in the Swiss-Prot comment line. In this case, the Swiss-Prot disease term with an OMIM reference was different from the extracted one. In the other cases, the discrepancy was due to an OMIM synonym (alternative title) classified into a distinct descriptor in MeSH. When this happened, a parent/child relationship was usually observed between the two MeSH terms mapped with either OMIM or Swiss-Prot diseases.

The performance of the mapping on ICD-10 was less good. This can be explained by the fact that, in contrast to the MeSH terminology, ICD-10 concepts are not enriched with synonyms

and lexical variants. We should refine the matching procedure to ICD-10 using more sophisticated natural language processing techniques such as normalisation and stemming.

Table 2: Mapping of 3197 disease comment lines of Swiss-Prot on MeSH and ICD-10 terms. SP∩OMIM means that both mappings correspond to the same MeSH descriptor.

	MeSH			ICD-10		
	Exact match	Partial match	Total	Exact match	Partial match	Total
SP	577 (18%)	819 (26%)	1396 (44%)	13 (7%)	37 (19%)	50 (25%)
OMIM	655 (20%)	680 (21%)	1335 (42%)	14 (7%)	40 (20%)	54 (27%)
SP ∩ OMIM	354 (11%)	390 (12%)	929 (29%)	6 (3%)	18 (9%)	30 (15%)
SP ∪ OMIM	866 (27%)	860 (27%)	1726 (54%)	21 (11%)	48 (24%)	69 (35%)

SP: Swiss-Prot

4 Discussion

In this study, we designed a mapping procedure to link the UniProtKB/Swiss-Prot human protein entries and the corresponding OMIM entries to the MeSH and ICD-10 disease terminologies. The procedure which combined exact and partial matches of disease names was able to map with good precision more than the half of the disease comment lines in UniProtKB/Swiss-Prot. Although this recall could be considered as satisfactory, we are currently trying to improve the coverage of the mapping.

The main problem encountered in the mapping process lay in the difference of granularity between the terminologies. MeSH is indeed relatively coarse-grained in terms of genetic diseases, and the situation with ICD-10 is even worse. Even if less specific terms exist, their matching scores are usually below the score threshold. For example, the disease term *recurrent intrahepatic cholestasis of pregnancy* matches with *intrahepatic cholestasis* with a score of -0.36, which is below the threshold. In fact, as the score threshold was set up from few data, it does not correspond to a clear cut in the recall/precision curve. Therefore, future improvement will consist in tuning the scoring schema to recover these terms. First, we can improve the word weighting by considering a common vocabulary resource for the word frequency calculation, and by excluding insignificant common words with a list of stopwords. Previous studies had shown the efficiency of methods using natural language processing pre-treatment, such as word normalisation or stemming, in terminology mapping processes [16,17]. We did not use such methods because the MeSH terms include various orthographic and lexical variants. However, a word normalisation step could have help mapped the disease *polycystic ovarian syndrome* to the MeSH term *polycystic ovary syndrome*. Second, it may be worth investigating ways to better exploit the hierarchical structure of these terminologies. For instance, a term such as *hypophosphatasia, adult type*, which matches with *hypophosphatasia* below the score, is clearly a child of this term. We should find ways to include this information in the score calculation. Such an attempt has been made, for instance, to categorise OMIM phenotypes using MeSH terms [18].

Nevertheless, the problem of MeSH granularity will hardly be completely solved by these methods. We need definitely to explore the mapping to other medical terminology resources, in particular the UMLS which provides other terminologies, together with relationships between concepts and semantic categories. This information could help identify related

concepts if an exact match is not available. Another direction will be to use the cited literature. Indeed, both UniProtKB and OMIM contain an important set of PubMed citations which are annotated with MeSH terms. Combining the similarity scores and term frequency in MEDLINE annotations will probably increase the chance of finding the correct term. We are currently working on this strategy. However, the best solution would be to complete MeSH and ICD-10 with new terms for genetic diseases. Our mapping could possibly help identify the gaps in these terminologies.

In conclusion, it becomes obvious that the use of a common terminology is required to help the integration of molecular biology data at clinical level. The indexing of human protein entries in UniProtKB with widely used disease terminologies will permit either clinicians or researchers to navigate from diseases to genes and from genes to diseases in an efficient way. Moreover, the ontological organisation of these terminologies will provide high-level search functionalities, such as the possibility to retrieve all genes involved in a class of diseases for a specific organ. This could be a major contribution in enforcing the interaction between biomedical researchers and clinicians for the benefit of research and patient care.

5 Acknowledgements

This work was funded by the Swiss National Science Foundation (grant No 3100A0-113970). We thank Robert Baud for help in ICD-10 resource handling.

6 References

[1] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res., 34**:** D322-D326, 2006.

[2] M. Ashburner, C.J. Mungall, and S.E. Lewis. Ontologies for biologists: a community model for the annotation of genomic data. Proc. Cold Spring Harbor Symp. Quant. Biol., 227–236, 2003.

[3] International Statistical Classification of Diseases and Health Related Problems (The) ICD-10, Second Edition. WHO Press, Geneva.

[4] K. Donnelly, “SNOMED-CT: The advanced terminology and coding system for eHealth”, Stud Health Technol Inform. 121:79-90, 2006.

[5] S.J. Nelson, M. Schopen, A.G. Savage, J.L. Schulman, and N. Arluk. The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. In: Proceedings of the 11th World Congress on Medical Informatics, San Francisco, Fieschi, M. et al., editors, CA. Amsterdam: IOS Press;. pp.67-69, 2004.

[6] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., 32: D267-D270, 2004.

[7] A.J. Butte and I.S. Kohane. Creation and implications of a phenome-genome network. Nat Biotechnol. 24(1):55-62, 2006.

[8] M.N. Cantor, I.N. Sarkar, O. Bodenreider, and Y.A. Lussier. GenesTrace: Phenomic knowledge discovery via structured terminology. Pac. Symp. Biocomput., pp.103-114, 2005.

[9] G. Marquet, A. Burgun, F. Moussouni, E. Guerin, F. Le Duff, and O. Loreal. BioMeKe: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis. Stud. Health Technol. Inform. 95:80-5, 2003.

[10] Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput., pp.64-75 2006.

[11] J. LaBaer. Mining the literature and large datasets. Nat Biotechnol., 21(9):976-977, 2003.

[12] M. Masseroli, O. Galati and F. Pinciroli. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res., 33:W717-W723, 2005.

[13] The UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res., 35: D193-D197, 2007.

[14] 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, 33:D514-517, 2005

[15] H. Shatkay. Hairpins in bookstacks: Information retrieval from biomedical text. Brief. Bioinform., 6:222-38, 2005.

[16] I.N. Sarkar, M.N. Cantor, R. Gelman, F. Hartel, and Y.A. Lussier. Linking biomedical language information and knowledge resources: GO and UMLS. Pac. Symp. Biocomput., pp.439-450., 2003.

[17] H.L. Johnson, K.B. Cohen, W.A. Baumgartner, Z. Lu, M. Bada. T. Kester, H. Kim, and L. Hunter. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. Pac. Symp. Biocomput., pp.28-39, 2006.

[18] M.A. van Driel, J. Bruggeman, G. Vriend, H.G. Brunner, and J.A. Leunissen. A text-mining analysis of the human phenome. Eur J Hum Genet., 14: 535-42, 2006.