Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics

Olivier Bodenreider, M.D., Ph.D.[a], Joyce A. Mitchell, Ph.D,[a,b] , Alexa T. McCray, Ph.D[a]
[a] National Library of Medicine, Bethesda, Maryland
[b] University of Missouri, Columbia
{olivier|mitchell|mccray}@nlm.nih.gov

***Objectives**: Terminology and knowledge resources are essential components of interoperability among disparate systems. This paper evaluates whether names and relationships needed in biomedical informatics are present in the UMLS. **Methods**: Terms for five broad categories of concepts were extracted from LocusLink and mapped to the UMLS Metathesausus. Relationships between gene products and the other four categories (phenotype, molecular function, biological process, and cellular component) were searched for in the Metathesaurus. All gene products in the Gene Ontology database were also mapped to the UMLS in order to evaluate its global coverage of the domain. **Results**: The coverage of concepts ranged from 2% (gene product symbols) to 44% (molecular functions). The coverage of relationships ranged from 60% for Gene product-Biological process to 83% for Gene product-Molecular function. **Discussion**: Terminology and ontology issues are discussed, as well as the need for integrating additional resources to the UMLS.*

INTRODUCTION

Integrating complex data, dynamic in nature, from heterogeneous resources, and without broadly applied standards constitutes a real challenge for users trying to make sense of the increasing amount of information made publicly available by a number of centers in the domain of bioinformatics and biomedical informatics [1]. Earlier studies have explored the field of ontology for molecular biology and bioinformatics [2-4]. Although not an ontology, the Unified Medical Language System® (UMLS®) appears as a possible candidate for supporting interoperability among available resources because of its broad coverage of the biomedical domain, including some coverage of bioinformatics. Instead of extending some of its components, as suggested by Yu and al. [5], we propose to evaluate the UMLS as a terminology and knowledge resource for biomedical informatics by exploring its coverage of terms and relationships needed for bioinformatics applications.

MATERIALS

UMLS

The resource evaluated in this study is the Unified Medical Language System (UMLS), developed and maintained by the National Library of Medicine. The UMLS Metathesaurus[1] (13th edition, 2002AA) contains over 1.5 million unique English strings drawn from more than sixty medical vocabularies, and organized in some 775,000 concepts. While broadly covering the clinical subdomain of biomedicine (over 150,000 concepts are categorized as disorders or findings), the UMLS also represents many genes and gene products, especially those included as supplementary concepts in the Medical Subject Headings (MeSH). In the UMLS, each concept is categorized by means of semantic types (STs) from the Semantic Network. Although most concepts are assigned to one ST, chemical concepts such as proteins are usually assigned to both one ST characterizing their structure (e.g., *Amino Acid, Peptide, or Protein*) and one ST characterizing their function (e.g., *Enzyme*).

Gene Ontology

The Gene Ontology™ project[2] "seeks to provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organisms". Gene Ontology (GO) is developed and regularly updated by the Gene Ontology Consortium. The three subdomains of GO are molecular functions, biological processes, and cellular components. Each subdomain is organized as an independent hierarchy of concepts (called "terms" in GO). GO does not provide an ontology of genes or gene products, but rather serves as a controlled vocabulary for collaborating centers to annotate their databases. The GO database, however, integrates annotation files, providing a link between gene and gene products on the one hand and the three subdomains of GO.

[1] umlsks.nlm.nih.gov
[2] www.geneontology.org/

LocusLink

LocusLink[3] is a gene-centered resource developed and regularly updated by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. LocusLink "organizes information around genes to generate a central hub for accessing gene-specific information" for various species. In other words, LocusLink offers a single interface to access gene-related, curated information including the names of the gene, its products, the diseases resulting from its mutations, and its functions (represented with concepts from GO and other ontologies). In addition to the summary integrated on one page, more detailed information is available through the many links to external, specialized sites (e.g., gene sequence, gene variants, literature about this gene). Integrating disparate information, LocusLink provides a simple means to gather knowledge about specific genes and was therefore a useful entry point for this study.

METHODS

The methods can be summarized as follows. First, a list of terms relevant to biomedical informatics is established from authoritative sources. Phenotype, gene and gene product, molecular function, biological process and cellular component are the categories of concepts studied in this paper. Second, these terms are mapped to concepts from the UMLS Metathesaurus. Finally, the presence of the relationships recorded in the original resources is checked in the UMLS Metathesaurus.

Using the methods developed for mapping LocusLink terms to the UMLS, the coverage of genomics terminology in the UMLS is evaluated by mapping all terms found in the Gene Ontology database to the Metathesaurus.

Establishing the list of terms

We queried LocusLink on January 11, 2002 requesting genes associated with a human disease and whose sequence was established (Query: has_seq AND disease_known; Organism: Human). 1276 loci were retrieved and downloaded as a structured text file.

The fields extracted from the file resulting from the LocusLink search consist of genes and gene products, diseases, and concepts drawn from the three subdomains of Gene Ontology. All fields corresponding to genes or gene products, including official names, synonyms and symbols, are categorized as Gene / Gene product. The identifier of the locus (LOCUSID) is used to identify relationships among fields within a locus.

Additionally, we extracted from Gene Ontology all concepts (called "terms" in the GO parlance), with their preferred name and synonyms, excluding those marked as obsolete, as well as all gene products from various species annotated with GO terms, also present in the GO database.

The number of terms in each field of LocusLink and GO is given in Table 1.

Mapping terms to UMLS concepts

The terms extracted from LocusLink and Gene Ontology were mapped to the UMLS by first attempting an exact match between the input term and Metathesaurus concepts. If an exact match failed, normalization was then attempted. This process makes the input and target terms potentially compatible by eliminating such inessential differences as inflection, case and hyphen variation, as well as word order variation. Duplicate terms were removed from each set prior to mapping to the UMLS.

Moreover, the mapping is considered successful only if the concept mapped to is semantically compatible with the original term. We created a compatibility table associating each of the five categories above with semantic types (STs) in the UMLS Semantic Network. For a given category, a given semantic type can qualify the mapping to a category (e.g., the ST *Disease or Syndrome* for the category Phenotype), block the mapping (e.g., the ST *Plant* for the category Molecular function), or simply be neutral, which most STs are for most categories. This method takes advantage of prior categorization of both the original terms (by the field type) and the target concepts (by the semantic types) to prevent most irrelevant mappings from happening. Examples of mappings rejected for semantic incompatibility include *merlin* (a gene product) to *Falco colombarius* (a bird also called *merlin*), and *oxygen sensor* (a molecular function) to *oxygen sensors* (a medical device).

Checking the relations against the UMLS

After mapping terms from LocusLink and GO to the UMLS, each locus can be seen as five sets of UMLS concepts, one for each category. Some sets may be empty if no mapping was found or selected for any of the terms in this category. Conversely, some sets may contain more than one concept either because one term mapped to several semantically compatible concepts, or because several terms mapped to a UMLS concept.

For a given locus, concepts from the Gene / Gene product (G/GP) category are associated with all concepts from the other four categories, resulting in

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

four categories of relationships: G/GP-Phenotype, G/GP-Molecular function, G/GP-Biological process, and G/GP-Cellular component. A total of 7919 pairs of concepts was generated.

The following kinds of relationships were checked against the UMLS Metathesaurus: hierarchical (using the parent/child and broader/narrower relationships), associative (using the "other" relationships), and co-occurrence relationships.

For a given pair of concepts (C_1,C_2), associative and co-occurrence relationships were checked not only between C_1 and C_2, but also between the parents of C_1 and C_2, between C_1 and the parents of C_2, and between the parents of both concepts. The rationale for allowing this extended set of relationships to be checked is that, in most cases, LL or GO concepts are very fine-grained while UMLS relationships may be recorded at a higher level in the hierarchy. This happens almost systematically with gene products, often found as supplementary concepts (SCs) in MeSH, while the co-occurrence relationships are recorded among descriptors, the SCs being descendants of the descriptors.

Hierarchical (and hierarchically-derived) relationships searched for include first- and second-generation ancestor and descendant, sibling (common first-generation ancestor) and cousin (common second-generation ancestor).

Finally, more than one relationship may be found between two concepts, either within a kind of relationships (e.g., parent and sibling), or across kinds (e.g., associative and co-occurrence relationships).

RESULTS

Results are expressed in terms of coverage, i.e., percentage of items from original sources represented in the UMLS. The result of the mapping process is "selected" when a term was mapped to the UMLS and semantic compatibility was assessed between the field type in the source and the semantic type of the concept mapped to, "rejected" when semantic compatibility could not be assessed, or "none" when a term failed to be mapped to the UMLS after normalization.

Coverage of concepts

The coverage of concepts is summarized in the right part of Table 1. The mapping rates ranged from 2 to 44%. Except for the class Biological process, names from the ontology part of GO had the best mapping rates. Disease names from LocusLink are also among the best mapping rates. Not surprisingly, the rate of mapping for the names and symbols of gene products was significantly lower, especially in GO with many non-human gene products. Rejected mappings were generally few, except for the symbols.

Coverage of LocusLink relationships

The coverage of relationships of gene and gene products (G/GP) to concepts from the other four categories studied is summarized in Table 2. Out of the 4255 pairs of concepts related in LocusLink after mapping to the UMLS, 2996 (70%) were also found related in the Metathesaurus. The percentage of relationships found varies across categories of pairs, from 60% for G/GP-Biological process to 83% for G/GP-Molecular functions. While the relationships of G/GP to molecular functions are often represented by both hierarchical and co-occurrence relationships, co-occurrences play a major role in the representation of the relationships of G/GP to the other four categories. Associative relationships are the second major source of relationships of G/GP to Phenotype and Molecular function. Finally, molecular functions are also often represented through hierarchical relationships, i.e., linked to an ancestor representing a functional category.

Finally, for 3664 pairs of terms associated in LocusLink, at least one of the terms failed to be mapped to the UMLS. Therefore, the relationship could not be studied in these cases.

EXAMPLE

Dystrophin (LocusLink ID: 1756) will be used to illustrate our method, and some of the difficulties we encountered along the way.

Dystrophin is the commonly used name for the eighteen gene products of the gene officially named "*dystrophin (muscular dystrophy, Duchenne and Becker types), includes DXS143, DXS 164, DXS206, DDXS 230, DXS239, DXS268, DXS269, DXS270, DXS272*". This official name comes from the HUGO Gene Nomenclature Committee[4] and is used in LocusLink, but not in the annotations in GO or the UMLS where the simpler *dystrophin* is used instead. Therefore, the automatic mapping of the gene name from LocusLink to the UMLS by the methods presented above would have resulted in a failure, leading to the false conclusion that *dystrophin* is not represented in the UMLS. This gene, when mutated, causes Duchenne and Becker muscular dystrophy as well as X-linked cardiomyopathy.

In the UMLS, *dystrophin* is hierarchically related to four first-generation ancestors: *muscle proteins*, *cytoskeletal proteins*, *actin-binding protein*, and *membrane proteins*. In addition, *dystrophin* has associative relationships to the disease concepts

[4] www.gene.ucl.ac.uk/nomenclature/

muscular dystrophies and *muscular dystrophy, Duchenne*. Two diseases names present in LocusLink are mapped to the latter concept: *Duchenne muscular dystrophy* and *Cardiomyopathy, dilated, X-linked.* However, although mapped to a concept, no direct association to *dystrophin* is found in the UMLS for the disease *Becker muscular dystrophy* mentioned in LocusLink. The two types of muscular dystrophies are associated in the UMLS, though. *Dystrophin* was recorded as a major descriptor in 3338 MEDLINE® citations over the last ten years, co-occurring with as many as 2615 different other descriptors, often with a low frequency (i.e., once or twice). Frequently co-occurring concepts include several variants of *muscular dystrophy* (Freq=654) and *muscle* (Freq=254), but also *cytoskeletal protein* (Freq=190), *gene therapy* (Freq=46) and *mutation* (Freq=43).

In GO, *dystrophin* is annotated with eight concepts. Three of these concepts map to UMLS concepts in the Metathesaurus and are found in co-occurrence relationship with *dystrophin*: one cellular component (*cytoskeleton*) and two biological processes (*muscle development* and *muscle contraction*). However, although not resulting in direct mappings, a strong correspondence can be found between three other GO concepts and first-generation ancestors of *dystrophin* in the UMLS: between the cellular component *extrinsic plasma membrane protein* and *membrane proteins* in the UMLS, between the molecular function *structural protein of cytoskeleton* and *cytoskeletal proteins* in the UMLS, and between the molecular function *actin binding* and the UMLS concept *actin-binding protein.* Along the same lines, *calcium binding protein*, co-occurring with *dystrophin* in the UMLS, can be seen as corresponding to the function *calcium binding* in GO. The representation of functions by either an associative relation to a function (GO) or a hierarchical relationship (UMLS) is studied in [6]. Finally, *dystrophin* is annotated by one other GO concepts for which there is no correspondence in the UMLS: *cell shape and cell size control.*

In general, gene products in GO tend to be annotated with concepts more specific than those to which they are related in the UMLS. Moreover, the structure of GO provides a simpler means of relating *dystrophin* to developmental processes such as *embryogenesis or morphogenesis* while it would require exploring the ancestors of co-occurring concepts in the UMLS to acquire this information. On the other hand, the UMLS shows *dystrophin* as the parent concept to supplementary concepts that correspond to some of its variant protein forms. GO does not have these concepts represented, nor does it carry the individual isoforms as separate concepts.

DISCUSSION

Terminology issues

Despite initiatives such as the International Protein Index[5] (IPI) and the HUGO Gene Nomenclature Committee mentioned earlier, naming the things needed to describe the world of bioinformatics remains a challenge, especially in the domain of genomics and proteomics. While some biomedical disciplines have developed and refined naming conventions over for several decades or centuries of existence, bioinformatics terminology is still in its infancy. Besides supporting reasoning, one role played by Gene Ontology is that of a controlled vocabulary, i.e., to promote a standard way of naming the concepts involved in bioinformatics, as reflected by the many databases using it to annotate gene products. Practically, in a near future, the availability of resources allowing to map names from one system to another are probably more desirable (and realistic) than the hypothetical enforcement of naming conventions. The principal reason why consistency in naming is important is because a lack of it certainly prevents useful resources (or, at least, useful bits of information) from being accessed. The example of *dystrophin* above offers an illustration of this issue.

Ontology issues

Some of the concepts represented in GO and LocusLink are logically, or systematically, polysemous [7]. Systematic polysemy is distinct from ambiguity. Ambiguous terms are homonyms that have entirely different meanings (e.g., *ventilation* is ambiguous since it has at least two different meanings, one referring to the environmental flow of air, and the other referring to respiration, a biological phenomenon). Systematically polysemous terms, on the other hand, represent meanings that are closely related to each other in predictable ways within a domain. The term itself is underspecified until it appears in a defined context. This phenomenon has implications for the design of an ontology, since these underspecified meanings need to be represented appropriately.

A classical example of systematic polysemy in bioinformatics is represented by the frequent use of the a unique name to stand for both a gene (i.e., the code) and a gene product (i.e., what is coded for). In this case, not only the two objects in the world are strongly associated from a cognitive standpoint, but also their nature is very close. Similarly, this domain offers many examples of systematic association between a gene and the disease caused by its mutation

[5] www.ebi.ac.uk/IPI/

(e.g., the gene called "*Wiskott-Aldrich syndrome (eczema-thrombocytopenia)*") or a gene product and a function (e.g., *galactokinase*) to name a few.

Interoperability vs. integration

Integrating all gene and gene product names in the UMLS would certainly make it much larger, but not necessarily more useful. In fact, including strings such as gene symbols would also increase ambiguity.

More importantly, what is really needed is better access to information, through, for example, cross-references, as well as better methods for detecting similarity among closely related concepts.

In conclusion, although somewhat limited in its coverage of the bioinformatics domain, especially for the finest-grained concepts, the UMLS represents many of the relationships found in LocusLink. Improved methods for mapping bioinformatics concepts to the UMLS based, for example, on systematic polysemy would also make it more useful.

References

1. Mitchell JA, McCray AT, Bodenreider O. From phenotype to genotype: Experiences in navigating the available information resources. Proc AMIA Symp 2002:(in press).
2. Creating the gene ontology resource: design and implementation. Genome Res 2001;11(8):1425-33.
3. Schulze-Kremer S. Ontologies for molecular biology. Pac Symp Biocomput 1998:695-706.
4. Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Brief Bioinform 2000;1(4):398-414.
5. Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Symp 1999:181-5.
6. Burgun A, Bodenreider O, Le Duff F, Mounssouni F, Loréal O. Representation of roles in biomedical ontologies: a case study in functional genomics. Proc AMIA Symp 2002:(in press).
7. Pustejovsky J. Type coercion and lexical selection. In: Pustejovsky J, editor. Semantics and the Lexicon. Dordrecht: Kluwer Academic Publishers; 1993. p. 73-94.

	Category	Field name	Number of names	Mapping to the UMLS (unique names)						
				selected	%	rejected	%	none	%	total
LocusLink	Phenotype	PHENOTYPE	1,893	575	34	5	0	1,122	66	1,702
	Gene / Gene product	OFFICIAL_GENE_NAME	1,244	244	20	18	1	982	79	1,244
		OFFICIAL_SYMBOL	1,244	200	16	39	3	1,005	81	1,244
		ALIAS_SYMBOL	2,743	394	15	175	7	2,100	79	2,669
		PRODUCT	1,502	266	18	9	1	1,185	81	1,460
		ALIAS_PROT	1,452	339	24	24	2	1,062	75	1,425
Gene Ontology	M. function	molecular function	5,626	2,436	44	13	0	3,136	56	5,585
	B. process	biological process	4,677	256	5	10	0	4,406	94	4,672
	C. component	cellular component	1,077	370	35	14	1	683	64	1,067
	Gene / Gene product	full_name	42,661	4,392	11	56	0	34,384	89	38,832
		symbol	62,366	1,167	2	439	1	58,775	97	60,381
		synonym	36,044	1,964	6	438	1	33,031	93	35,433
Total			162,529	12,603	8	1,240	1	141,871	91	155,714

Table 1 - LocusLink and Gene Ontology fields with their categorization (letf) and mapping to the UMLS (right).

Categ. 1	Category 2	At least one rel.		No relationships found		Total	Assoc.	Co-oc.	Hier.
Gene / Gene product	Phenotype	644	62%	387	38%	1,031	231	621	4
	M. function	1,022	83%	208	17%	1,230	187	922	788
	B. process	637	60%	421	40%	1,058	48	618	120
	C. component	693	74%	243	26%	936	65	680	66
Total		2,996	70%	1,259	30%	4,255	531	2,841	978

Table 2 - Type of LocusLink relationship found in the UMLS by category.