Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: A feasibility study

Olivier Bodenreider[a], Anita Burgun[b], Joyce A. Mitchell[a,c]

[a] *US National Library of Medicine, National Institutes of Health, Dept. of Health & Human Services, USA*
[b] *University of Rennes I, France*
[c] *University of Missouri, Columbia, USA*

Abstract

Objectives: While several sources of biomedical knowledge are available, these resources are often highly specialized and usually not suitable for a lay audience. This paper evaluates whether concepts needed for molecular biology and genetic diseases are present in WordNet, the electronic lexical database. **Methods**: Terms for four broad categories of concepts (phenotype, molecular function, biological process, and cellular component) were extracted from LocusLink and mapped to WordNet. All terms from the Gene Ontology database (gene products and ontology concepts) were also mapped to WordNet in order to evaluate its global coverage of the domain. Additionally, we tested two methods for improving the mapping of genetic disease names to WordNet. **Results**: The coverage of concepts ranged from 0% (gene product symbols) to 2.8% (cellular components). Removing specialization markers from the terms and using synonyms significantly increased the rate of mapping of genetic disease names to WordNet. **Conclusions**: Many of the most common single gene disorders are present in WordNet, as well as many high-level concepts in Gene Ontology. Therefore, WordNet is likely to be a useful source of lay knowledge in the framework of a consumer health information system on genetic diseases.

Keywords:

WordNet; Genetic diseases; Molecular biology; Consumer health information.

1. Introduction

Many repositories of information about genetic diseases and molecular biology have been created and most of them are publicly available. Examples of such resources include OMIM[1], GenBank[2], Swiss-Prot[3], and GeneCards[4]. Hubs such as LocusLink[5] regroup for a given gene many of the information scattered in these disparate and heterogeneous databases. While these resources are generally very useful to researchers, they are often difficult to use for a lay audience. Part of the difficulty of providing information to the general public lies in the high degree of specialization of most resources [1]. While simpler resources must be developed, an alternative approach consists of trying to make existing resources easier to understand by linking specialized terms to their definition (textual or formal, i.e., through the relationships of a given term to other terms).

Several hundreds of papers report uses of WordNet for various tasks[6] including machine translation, word sense disambiguation, text understanding, and question answering. Often, these tasks are conducted on general corpora such as the World Wide Web or newswires. Few studies, however, were published on using WordNet for education [2] or in specialized domains [3]. In earlier papers, we reported on the mapping of specialized terms from the Unified Medical Language System® (UMLS®) to WordNet [4] and on the properties of the definitions

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM (all URLs valid as of November 16, 2002)
[2] http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
[3] http://www.expasy.ch/sprot/
[4] http://bioinfo.weizmann.ac.il/cards/index.html
[5] http://www.ncbi.nlm.nih.gov/LocusLink/
[6] http://www.seas.smu.edu/~rada/wnb/

Stud Health Technol Inform 2003;95:379-384.

of anatomical terms in WordNet [5]. These two studies encouraged us to evaluate WordNet as a possible source of lay knowledge for molecular biology and genetic diseases.

As a preliminary study, testing the feasibility of this broader endeavor, we examine how much of the terminology required is represented in WordNet. We use the terms present in LocusLink and the Gene Ontology as a surrogate for the required terminology.

2. Materials

WordNet

WordNet® [7] is an electronic "lexical database for the English language", developed and maintained at Princeton University since 1985. Although not specialized in any particular subdomain, WordNet contains, as the English language does, many terms used in the biomedical domain. Sets of synonymous terms, or synsets, constitute its basic organization. The current version (1.7.1) integrates over 110,000 synsets organized in separate hierarchies for nouns, verbs, adjectives, and adverbs. Several types of relations between synsets are recorded in WordNet, including hyponymy (specific-generic) and meronymy (part-whole) among nouns.

Gene Ontology

The Gene Ontology™ project[8] "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 of genes and gene products. 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 on the other.

LocusLink

LocusLink[9] 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.

Differences among the sources

While WordNet is not expected to contain many gene names, many disease names, including names of genetic diseases, are part of the English language and are represented in WordNet, e.g., *Huntington's disease* (11943647). However, genetic disease names from LocusLink sometimes fail to map to WordNet, not because the concept for the disease is not represented in WordNet, but because the names for the disease are different. For example, the genitive marker ('s) found in the term in WordNet prevents the LocusLink term *Huntington disease* from mapping to WordNet. Another characteristic exhibited by many LocusLink disease names is their complexity. Examples of complex names include *Porphyria, acute intermittent* and *Muscular dystrophy, Duchenne-like, type 2*. Neither term exists in WordNet, but, in both cases, a more general concept does (*porphyria, muscular dystrophy*).

[7] http://www.cogsci.princeton.edu/~wn/
[8] http://www.geneontology.org
[9] http://www.ncbi.nlm.nih.gov/LocusLink/

3. Methods

The methods can be summarized as follows. First, a list of terms relevant to molecular biology and genetic diseases is established from LocusLink and Gene Ontology. 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 synsets from the noun hierarchy of WordNet. Then, each mapping is validated or rejected, based on semantic constraints, i.e., the comparison between the category of the original term and the category computed for WordNet synsets from hierarchical information. In addition to studying the coverage of molecular biology and genetic diseases in WordNet, we later focus on genetic disease names found in LocusLink and examine various methods to improve the mapping of these terms to WordNet.

Establishing the list of terms

We queried LocusLink on October 11, 2002 requesting genes associated with a human disease and whose sequence was established (Query: has_seq AND disease_known; Organism: Human). 1371 loci were retrieved and downloaded as a structured text file. From this file, we extracted the fields containing genes, gene products, and diseases. All fields corresponding to genes or gene products, including official names, synonyms and symbols, were categorized as Gene / Gene product. The identifier of the locus (LOCUSID) was used to identify relationships among fields within a locus.

Additionally, we extracted from Gene Ontology (GO) 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 (left part). Duplicate terms were removed from each set prior to mapping to WordNet.

Mapping terms to WordNet synsets

The terms extracted from LocusLink and Gene Ontology were mapped to WordNet using *wn*, a program provided as part of the WordNet distribution. Input terms processed through this interface are normalized by the WordNet "morphology processor", Morphy, using a set of rules and a list of exceptions. Normalization makes the input and target terms potentially compatible by eliminating such inessential differences as inflection, case and hyphen variation[10].

For each term from LocusLink and Gene Ontology, the presence of an equivalent lexical item in WordNet was recorded. One term may map to several noun synsets in WordNet. For example, *cell* maps to six distinct WordNet synsets including *prison cell* and *electric cell*.

Computing a semantic category for WordNet synsets

In WordNet, a semantic class can be defined as the set of all hyponyms of a given synset (i.e., all direct hyponyms and their hyponyms, all the way down). For example, the hyponyms of *disease* (11865979) constitute the semantic class Disease. Conversely, any of the synsets in the class Disease can be assigned the category Disease. In practice, a broad class may be defined as the union of several classes not necessarily overlapping. For example, the class Gene / Gene Product results from the union of classes for Gene and Gene product.

We established classes of WordNet noun synsets corresponding to the five categories of concepts studied in this paper (Phenotype, Gene / Gene product, Molecular function, Biological process, and Cellular component) by seeding the classes manually with relevant high-level noun synsets. We then populated the first four classes by adding to the class all the hyponyms of each seed synset. For example, the class Biological process is seeded with the two synsets *biological process* (11410462) and *organic phenomenon* (09385337) and includes *catabolism* (11343680), *apoptosis* (09445312) and *cell-mediated immune response* (00650432).

[10] Contrary to *norm*, the lexical program used for normalizing terms in the Unified Medical Language System (UMLS), Morphy does not ignore word order and the genitive marker.

The technique used for the class Cellular component was slightly different because the relationship of a cellular component to cell is meronymy (part-whole), not hyponymy (specific-generic). Therefore, the class Cellular component was seeded with the synset *cell* (00004239) and populated with both the meronyms of *cell* hyponyms and the hyponyms of *cell* meronyms. Here again, hyponyms and meronyms are computed all the way down. For example, the class Cellular component includes *acrosome* (04684711), which is a part of *sperm cell* – itself a kind of *cell*, and *mitochondrion* (04672846), which is kind of *organelle* – itself a part of the *cell*.

The list of noun synsets used as seeds for the five categories is given in Table 2.

Validating mappings to WordNet with semantic constraints

In order to validate the mapping of terms from LocusLink and Gene Ontology to WordNet synsets, we compared the category of the original term (C_o) to the category computed for the synset mapped to (C_w) using the method described in the previous section. The mapping is declared valid when the semantic constraint is satisfied, i.e., when C_o and C_w are the same, and rejected otherwise. For example, the term *REVERSE TRANSCRIPTASE* is mapped to the noun synset *reverse transcriptase* (12670183). The original term is found in LocusLink as 'FULL_NAME', i.e., by design, a member of the class Gene / Gene product (see Table 2). The mapping is deemed valid because this synset has *activator* as one of its hypernyms and, therefore, belongs to the class Gene / Gene product, as the original term does. In contrast, the mapping of the gene symbol *da* (for the gene *daughterless* in Drosophila melanogaster) to the synset "*district attorney, DA*" (08224010) is rejected because their categories are different.

Terms may map to more than one synset in WordNet. However, applying the semantic constraints usually results in mapping to only one relevant synset. For example, *cell* has six senses in WordNet, only one of which (sense 2, 00004239) corresponds to biology and is semantically compatible with Cellular component.

Sometimes, more than one synset mapped to is semantically compatible with the original term. In this case, all compatible synsets are recorded and the mapping is considered multiple. Examples of multiple mappings include *reproduction* to "the process of generating offspring" (11429956) and "the sexual activity of conceiving and bearing offspring" (00641241). Although only the former corresponds to the biological process found in Gene Ontology ("The production by an organism of new individuals that contain some portion of their genetic material inherited from that organism"), both mappings are deemed valid because the synset *biological process* (11410462) is a hypernym of both synsets, making them semantically compatible with the class Biological process.

Improving the mapping of genetic disease names to WordNet

In order to compensate for the differences mentioned earlier between LocusLink and WordNet, we used two different approaches and applied them first separately and then in combination. The first approach consisted in simplifying disease terms from LocusLink by removing from them any mention of specialization (e.g., "type I"), of etiology (e.g., "due to PTS deficiency"), and of accompanying disease or symptom (e.g., "with bilateral retinoblastoma"). More generally, we removed the part of the term trailing the leftmost comma, if any (e.g., extracting *Meningioma* from *Meningioma, NF2-related, sporadic*). The second method consisted in augmenting LocusLink disease names with synonyms from the Unified Medical Language System[11] (UMLS) Metathesaurus®, when the original term could be mapped the UMLS. A Metathesaurus concept is a cluster of "synonyms" corresponding to a meaning, including synonymous terms (e.g., *Huntington's Disease*, *Huntington's Chorea*) as well as lexical variants (e.g., *Huntington's Disease*, *Huntington Disease*). Each synonym was then mapped to WordNet in addition to the original term, increasing the chances for finding a mapping. Finally, we combined the two approaches, first simplifying the original term and then mapping the simpler term to the UMLS to acquire synonyms. Only the combination of approaches allowed, for example, *Wilms tumor, type 1* to map to the WordNet synset *Wilms' tumor* (12025266).

[11] http://umlsinfo.nlm.nih.gov

4. Results

Coverage of molecular biology and genetic disease concepts in WordNet

The coverage of concepts is summarized in the right part of Table 1. The mapping rates are generally very low, ranging from 0 to 2.8%. Disease names from LocusLink and cellular components from Gene Ontology are among the best mapping rates. Not surprisingly, the rate of mapping for the names and symbols of gene products was even lower. Rejected mappings were generally few, except for the symbols (often mapped to acronyms in WordNet) and for cellular components (our validation process, by design, essentially restricted cellular components to the structural components of the cell, while Gene Ontology also includes macromolecular complexes present in the cell).

Improvement in the mapping of genetic disease names to WordNet

As shown in Table 3, only 47 (2.5%) of the 1903 genetic disease names from LocusLink mapped to a synset in WordNet without transformation. Simplifying terms resulted in 473 new mappings (+24.9%), boosting the overall mapping rate to 27.4%. However, after simplification, 520 original disease names mapped to only 175 distinct synsets, suggesting that the simplified term might be a hypermym of the original term (e.g., *Muscular dystrophy, limb-girdle, type 1A* and *Muscular dystrophy, Duchenne-like, type 2* both map to *muscular dystrophy* after simplification). Augmenting the original terms with synonyms from the UMLS Metathesaurus led to a smaller increase (51 additional mappings, +2.7%), but accounted for the mapping of many eponymic disease names featuring no genitive marker (e.g., *Huntington disease*) that would not have been mapped otherwise. Finally, the best result came from combining the two methods. With 550 new mappings (+28.9%), the number of new mappings resulting from the combined approaches was slightly better than the sum of the number of new mappings obtained with each method when applied in isolation. The best mapping rate overall is 31.4%.

5. Future work and conclusions

The goal of this study was to obtain a quantitative evaluation of the presence in WordNet of terms from the domain of molecular biology and genetic diseases. While an automatic mapping based on lexical resemblance and semantic constraints served this purpose well, a validation of the mapping by domain experts is required prior to using it in an application.

This study proved that many terms from this specialized domain are present in WordNet, including many of the most common single gene disorders (which are the most likely to be presented in a consumer health information system). These terms also correspond to high-level concepts in Gene Ontology (GO), which can be used as hooks for more specialized GO terms. For this reason, WordNet is likely to be a useful source of lay knowledge and definitions in the framework of a consumer health information system on genetic diseases.

When a term maps to WordNet, two types of definitions become available. A textual definition is provided in the gloss attached to the synset (e.g., *muscular dystrophy*: “any of several hereditary diseases of the muscular system characterized by weakness and wasting of skeletal muscles”). A formal definition can be derived from the position of the synset in the hierarchy of hyponyms and meronyms. For example, *Gaucher’s disease* has two direct hypernyms: *lipidosis* (a descendant of *metabolic disorder*) and *monogenic disorder* (a descendant of *genetic disorder*). In addition, WordNet may also provide valuable lay synonyms, clustered together in the synset with the specialized term (e.g., *Lou Gehrig's disease* for *amyotrophic lateral sclerosis*). Next, we will study how this may help consumers understand genetic diseases.

6. References

[1] Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. *Health Educ Res* 2001;16(6):671-92.

[2] Kegl J. Machine-readable dictionaries and education. In: Walker DE, Zampolli A, Calzolari N, editors. *Automating the lexicon : research and practice in a multilingual environment*. Oxford ; New York: Oxford University Press; 1995. p. 249-284.

[3] Strzalkowski T, Brandow R. Spotting technical concepts in natural language text. In: Stewman JH, editor. *Proceedings of the 9th Florida Artificial Intelligence Research Symposium*; 1996. p. 66-70.

[4] Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. *Proceedings of the NAACL'2001 Workshop, "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"* 2001:77-82.

[5] Bodenreider O, Burgun A. Characterizing the definitions of anatomical concepts in WordNet and specialized sources. *Proceedings of the First Global WordNet Conference* 2002:223-230.

7. Address for correspondence

Olivier Bodenreider, NLM, 8600 Rockville Pike – MS 43, Bethesda, MD 20894, USA - Email: olivier@nlm.nih.gov

Table 1 — LocusLink and Gene Ontology fields with their categorization (left) and mapping to WordNet (right)

	Category	Field name	Number of names	Mapping to WordNet (unique names)						
				selected	%	rejected	%	none	%	total
LocusLink	Phenotype	PHENOTYPE	2,131	47	2.5	2	0.1	1,854	97.4	1,903
	Gene / Gene product	OFFICIAL_GENE_NAME	1,338	12	0.9	3	0.2	1,323	98.9	1,338
		OFFICIAL_SYMBOL	1,338	5	0.4	45	3.4	1,288	96.3	1,338
		ALIAS_SYMBOL	3,532	12	0.4	186	5.5	3,175	94.1	3,373
		PRODUCT	1,734	10	0.6	3	0.2	1,664	99.2	1,677
		ALIAS_PROT	1,792	8	0.5	6	0.3	1,750	99.2	1,764
Gene Ontology	M. function	molecular function	5,868	74	1.3	27	0.5	5,762	98.3	5,863
	B. process	biological process	5,340	74	1.4	46	0.9	5,216	97.8	5,336
	C. component	cellular component	1,238	34	2.8	36	2.9	1,152	94.3	1,222
	Gene / Gene product	full_name	46,901	55	0.1	32	0.1	40,475	99.8	40,562
		symbol	123,868	23	0.0	427	0.4	120,254	99.6	120,704
		synonym	36,538	21	0.1	359	1.0	35,493	98.9	35,873
Total			231,118	375	0.2	1,172	0.5	219,406	99.3	220,953

Table 2 — List of WordNet hypernyms used to seed the semantic classes

Semantic class	WordNet (noun synsets)
Phenotype	*abnormalcy* *defect* (sense 1) *disorder* (sense 1) *ill health* *mental illness* *mental retardation* *symptom*
Gene / Gene product	*activator* *antigen* *carcinogen* *gene* *hormone* *inhibitor* *lysin* *macromolecule* *metabolite* *nucleotide* *pigment* *substrate* (sense 1) *toxin*
Molecular function	(same as Gene / Gene product)
Biological process	*biological process* *organic phenomenon*
Cellular component	*cell* (sense 2)

Table 3 — Improvement in the mapping of genetic disease names to WordNet

	Mapping	Simplification		Synonyms		Sim. + Syn.	
No transformed term created	none	611	32.1%	1301	68.4%	283	14.9%
	rejected	2	0.1%	2	0.1%	2	0.1%
	selected	*47*	*2.5%*	*47*	*2.5%*	*47*	*2.5%*
At least one transformed term created	none	747	39.3%	502	26.4%	997	52.4%
	rejected	23	1.2%	0	0.0%	24	1.3%
	selected	*473*	*24.9%*	*51*	*2.7%*	*550*	*28.9%*
Total selected		*520*	*27.4%*	*98*	*5.2%*	*597*	*31.4%*