Nucleic Acids Research, 2004, Vol. 32, Database issue ***D267–D270***
DOI: 10.1093/nar/gkh061

The Unified Medical Language System (UMLS): integrating biomedical terminology

Olivier Bodenreider*

Lister Hill Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received August 7, 2003; Revised and Accepted September 27, 2003

ABSTRACT

The Unified Medical Language System (http://umlsks.nlm.nih.gov) is a repository of biomedical vocabularies developed by the US National Library of Medicine. The UMLS integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. Vocabularies integrated in the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM and the Digital Anatomist Symbolic Knowledge Base. UMLS concepts are not only inter-related, but may also be linked to external resources such as GenBank. In addition to data, the UMLS includes tools for customizing the Metathesaurus (MetamorphoSys), for generating lexical variants of concept names (lvg) and for extracting UMLS concepts from text (MetaMap). The UMLS knowledge sources are updated quarterly. All vocabularies are available at no fee for research purposes within an institution, but UMLS users are required to sign a license agreement. The UMLS knowledge sources are distributed on CD-ROM and by FTP.

INTRODUCTION

Biomedical resources available to researchers are plentiful, ranging from databases of gene and protein sequences (1) and databases organized around one model organism (2), to integrative resources such as LocusLink (3), as well as biomedical literature databases (4) and biomedical ontologies (5). One common denominator for all of these resources is terminology, i.e. the names of genes, proteins, diseases, molecular functions, etc., in biomedical texts and the corresponding entries in the various controlled vocabularies and nomenclatures associated with these resources [e.g. the Medical Subject Headings (MeSH) for the literature, names and symbols of genes approved by the HUGO Gene Nomenclature Committee]. However, having identified terminology as a key integrating factor for biomedical resources does not imply that all resources have adopted standard vocabularies, which—whenever existing—would make these resources interoperable. While annotation databases such as Swiss-Prot/TrEMBL have greatly benefited from the development of Gene Ontology, which provides a controlled vocabulary for annotating gene products across model organisms, terminology standardization is far less advanced in other domains such as gene and protein names.

Research projects such as TAMBIS have addressed the specific issue of integrating disparate resources for bioinformatics through a model of domain knowledge (6). While TAMBIS provides a framework for integrating resources, its coverage is currently limited to five sources. In this paper, we present a different approach to information integration through terminology integration: the Unified Medical Language System® (UMLS®), developed over more than 15 years, which covers the entire biomedical domain. The UMLS was developed by the National Library of Medicine (NLM) as 'an effort to overcome two significant barriers to effective retrieval of machine-readable information' (7): the variety of names used to express the same concept and the absence of a standard format for distributing terminologies. By integrating more than 60 families of biomedical vocabularies, the UMLS Metathesaurus® currently provides not only an extensive list of names (2.5 million) for its 900 551 concepts, but also over 12 million relations among these concepts. Its scope is broader and its granularity finer than that of any of its source vocabularies.

After a brief presentation of the terminological resources integrated in the UMLS Metathesaurus, we show through an example how the UMLS may be useful to bioinformaticists. We conclude by presenting UMLS-related tools.

TERMINOLOGICAL RESOURCES INTEGRATED IN THE UMLS METATHESAURUS

The major component of the UMLS is the Metathesaurus, a repository of inter-related biomedical concepts. The two other knowledge sources in the UMLS are the Semantic Network, providing high-level categories used to categorize every Metathesaurus concept, and lexical resources including the SPECIALIST lexicon and programs for generating the lexical variants of biomedical terms. The Metathesaurus is the only resource presented in detail in this paper. Unless otherwise specified, the version described here is 2003AB (July 2003). The UMLS knowledge sources are updated quarterly.

*Tel: +1 301 435 3246; Fax: +1 301 480 3035; Email: olivier@nlm.nih.gov

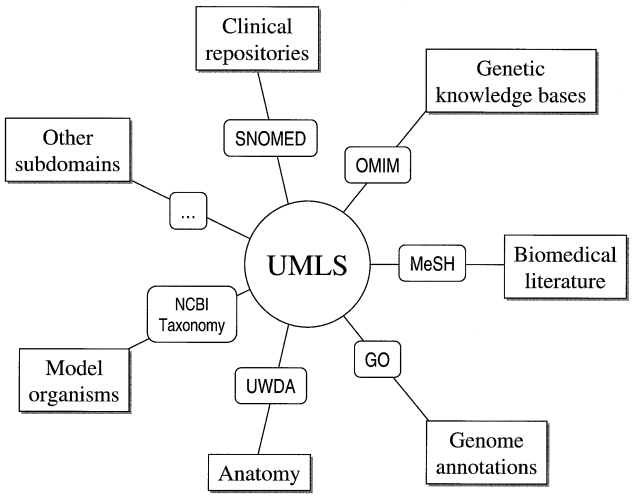

Figure 1. The various subdomains integrated in the UMLS.

Terminology of interest for bioinformaticists

Although the UMLS was not specifically developed for the needs of bioinformaticists, it includes terminologies used in bioinformatics. For example, recently integrated terminologies include the NCBI taxonomy, used for identifying organisms, and Gene Ontology, used for the annotation of gene products across various model organisms. The Metathesaurus also covers the biomedical literature with the MeSH, the controlled vocabulary used to index MEDLINE. Core subdomains such as anatomy, used across the spectrum of biomedical applications, are also represented in the Metathesaurus with the Digital Anatomist Symbolic Knowledge Base. Finally, the subdomain represented best is probably the clinical component of biomedicine, with general terminologies such as SNOMED® International (and soon SNOMED-CT®), Clinical Terms Version 3 and the International Classification of Diseases, to name a few. Clinical genetics resources include the Online Mendelian Inheritance in Man™ (OMIM™), represented in part, and the Online Multiple Congenital Anomaly/Mental Retardation (MCA/MR) Syndromes©. Other categories of terminologies in the Metathesaurus include specialized disciplines (e.g. nursing, psychiatry) and components of the clinical information system (e.g. diseases, drugs, procedures, adverse effects). Figure 1 illustrates how the UMLS Metathesaurus, by integrating these various terminologies, can serve as a link between not only the vocabularies, but also the subdomains they represent.

Terminology integration principles

In the UMLS, knowledge is organized by concept (i.e. meaning) (8). Synonymous terms are clustered together to form a concept and concepts are linked to other concepts by means of various types of relationships, resulting in a rich graph. Inter-concept relationships are either inherited from the structure of the source vocabularies or generated specifically by the editors of the Metathesaurus. Symbolic relationships can be hierarchical (e.g. 'is a kind of' or 'isa', 'part of') or associative (e.g. 'location of', 'caused by'). Statistical relations between concepts from the MeSH vocabulary are also present, derived from the co-occurrence of MeSH indexing terms in MEDLINE citations. Finally, each Metathesaurus concept is broadly categorized by means of the semantic types (i.e. the 135 high-level categories found in the Semantic Network), assigned by the Metathesaurus editors.

Such a structure makes it easy for users to perform tasks such as:

(i) collecting the various terms used to name a concept;

(ii) extracting the relations of one concept to other concepts, either hierarchical or associative, symbolic or statistical; and

(iii) obtaining a set of concepts for a given category, using the list of concepts that were assigned a given semantic type.

More formally, synonymy is the lexical relation used to cluster biomedical terms into concepts. Hyponymy ('isa') and meronymy ('part of') relations provide the hierarchical framework on which the concepts are organized. Associative relations, including co-occurrence relations, extend this framework laterally, providing links across various subdomains. The categorization by semantic type can be thought of as redundant with some of the hierarchical relations. In practice, because the categorization is independent of the structure of the source vocabularies, it provides a simple and stable means of semantic orientation in the Metathesaurus.

External cross-references

Biomedical terminologies often contain more information than the mere terms and their inter-relations. Beside definitions, additional information may include cross-references, either internal (e.g. 'See also…' in MeSH, treated as associative relations) or external (i.e. cross-references to other terminologies or databases). In most cases, this information is represented in the Metathesaurus. For example, MeSH supplementary concept records include many proteins for which a GenBank identifier is provided. Similarly, whenever it is relevant, concepts from the Online MCA/MR Syndromes point to the related diseases in MeSH and OMIM, even when the corresponding OMIM concept is not in the Metathesaurus. Examples of such cross-references are provided below.

EXTENDED EXAMPLE

Neurofibromatosis 2 is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features (9). The disorder is caused by mutations of the *NF2* gene resulting in the absence or inactivation of the protein product. The protein product of *NF2* is commonly called merlin (but also Neurofibromin 2 and Schwannomin) and functions as a tumor suppressor. Neurofibromatosis 2, *NF2* and Merlin are concepts in the UMLS, for which the Metathesaurus provides many synonyms, including those listed above. As shown in Figure 2, these three concepts are linked by associative relationships. Additionally, each concept is part of a hierarchy of concepts. Neurofibromatosis 2 inherits from ancestors such as 'Benign neoplasms of cranial nerves', which reflects the non-malignant behavior of schwannomas. Similarly, the function of *NF2* is expressed through its direct parent 'Tumor suppressor genes'. Semantic types from the UMLS semantic network provide a direct categorization to Metathesaurus concepts, making it easy to distinguish between the disease Neurofibromatosis 2 (Neoplastic Process) and the gene *NF2* (Gene or Genome).

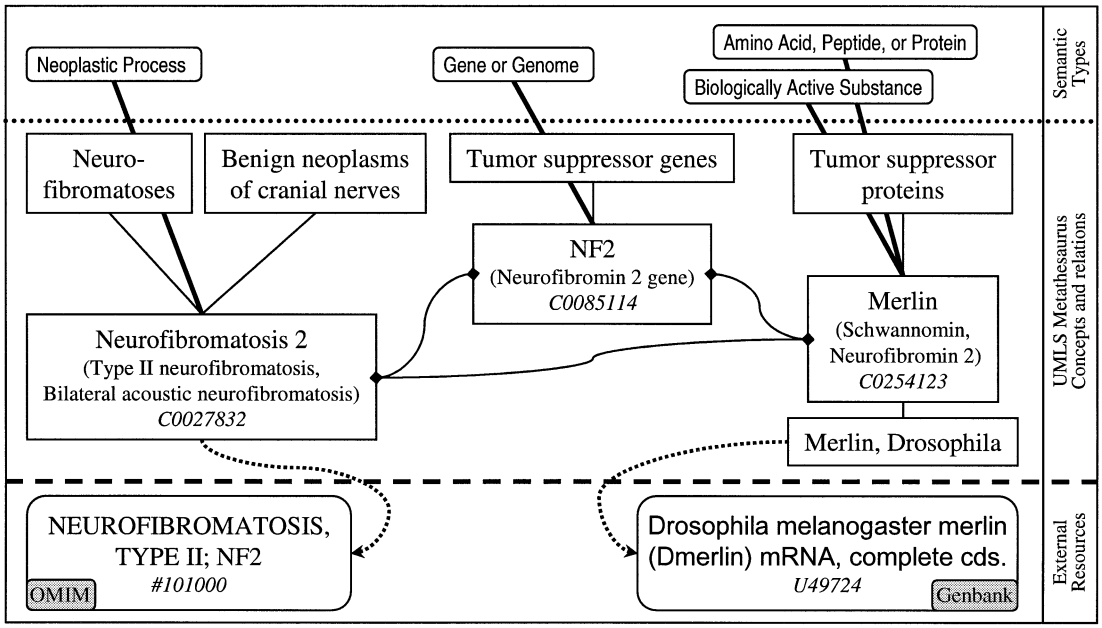

Figure 2. NF2 and related proteins and diseases in the UMLS (partial representation).

Among the descendants of Merlin, the concept 'Merlin, *Drosophila*' shows the existence of a homologous protein product in fruit fly. Similar relations are expressed through the co-occurrence of MeSH descriptors in MEDLINE citations. The three central concepts in our example tend to co-occur with each other with a high frequency, which is expected. More interestingly, concepts associated with them with a lesser frequency may reveal more recently established findings (e.g. the association between Merlin and the cytoskeleton). Finally, in addition to intra-Metathesaurus links, external cross-references allow UMLS users to bridge between terminologies in the UMLS and external resources. The OMIM identifier given for neurofibromatosis 2 provides a bridge to the OMIM database. Similarly, 'Merlin, *Drosophila*' can be searched in GenBank with the identifier provided by MeSH.

ACCESSING UMLS DATA

There is no fee associated with accessing the UMLS knowledge sources. However, UMLS users are required to sign a license agreement prior to accessing the data. Under this agreement, all vocabularies can be used for research purposes within an institution, but restrictions may apply for the use of some vocabularies in other kinds of applications. The UMLS knowledge sources are available on CD-ROM and by FTP.

The UMLS relational files can be loaded into a local database and accessed through SQL queries. Alternatively, an object-oriented model of the UMLS has been recently developed, allowing users to build Java applications against the NLM's repository through an application programming interface (API), as well as XML-based queries. The Knowledge Source Server is a web-based tool developed for the visualization and navigation of UMLS data. These three access mechanisms require users to be UMLS licensees.

UMLS RELATED TOOLS

In addition to data, the UMLS also consists of tools, either included as programs in the distribution or available as web-based services. MetamorphoSys helps users to customize the Metathesaurus for their applications by, e.g. selecting concepts from a given subdomain and selecting the preferred name of concepts. The program lvg, based on the SPECIALIST lexicon and hand-coded rules, allows users to generate lexical variants based on inflection (e.g. the plural form from the singular) and derivation (e.g. the adjectival form of a noun), as well as to perform other tasks useful for term mapping and information retrieval such as removing semantically unimportant words (called stop words) or abstracting away from hyphen variation. Both MetamorphoSys and lvg are part of the UMLS distribution. MetaMap (10), accessible as a web service, extracts Metathesaurus concepts from text. The input of MetaMap can be text of variable length and its output is a ranked list of Metathesaurus concepts associated with each piece of text. In addition to exact matches, MetaMap takes advantage of term variants generated by lvg and allows partial matches.

FOR FURTHER INFORMATION

In addition to the documentation distributed with the UMLS, there are more than 420 articles published on various aspects of UMLS properties and uses. The UMLSinfo website (http://umlsinfo.nlm.nih.gov) provides information and tools to users of the UMLS, including educational materials and answers to frequently asked questions. The umls-users mailing list also serves as a vehicle for news and support from NLM, as well as exchanges among users. Inquiries should be directed to custserv@nlm.nih.gov or 1-888-FINDNLM (+1 888 346 3656).

ACKNOWLEDGEMENTS

The author wishes to thank Stephanie M. Morrison and Natalie E. Krasikov, two resident clinical genetics specialists at NLM developing content for the *Genetics Home Reference* (http://ghr.nlm.nih.gov), for reviewing this manuscript and providing helpful suggestions.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
2. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
3. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
4. McEntyre,J. and Lipman,D. (2001) PubMed: bridging the information gap. *CMAJ*, **164**, 1317–1319.
5. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
6. Stevens,R., Baker,P., Bechhofer,S., Ng,G., Jacoby,A., Paton,N.W., Goble,C.A. and Brass,A. (2000) TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, **16**, 184–185.
7. Lindberg,D.A., Humphreys,B.L. and McCray,A.T. (1993) The Unified Medical Language System. *Methods Inf. Med.*, **32**, 281–291.
8. McCray,A.T. and Nelson,S.J. (1995) The representation of meaning in the UMLS. *Methods Inf. Med.*, **34**, 193–201.
9. Baser,M.E., Evans,R.D.G. and Gutmann,D.H. (2003) Neurofibromatosis 2. *Curr. Opin. Neurol.*, **16**, 27–33.
10. Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, 17–21.