Experiences in visualizing and navigating biomedical ontologies and knowledge bases

Olivier Bodenreider

U.S. National Library of Medicine

olivier@nlm.nih.gov

The biomedical domain is extensive and complex. From the perspective of information technology, this complexity translates into multiple, heterogeneous databases, offering limited interoperability. The first step towards getting a better understanding of these resources is for users to be able to visualize and navigate them. This paper presents some issues in visualizing and navigating biomedical ontologies and knowledge bases through case studies of two applications developed at the U.S. National Library of Medicine. SemNav[1] is a browser developed for visualizing and navigating biomedical concepts from the Unified Medical Language System® (UMLS®). GenNav[2], developed more recently, allows users to visualize the Gene Ontology™ (GO) database graphically and to navigate between concepts in GO, gene products annotated with these concepts, and the literature used as evidence for these annotations.

Background

UMLS

The Unified Medical Language System[3] (UMLS), developed and maintained by the U.S. National Library of Medicine, comprises two major components, the Metathesaurus® and the semantic network. The UMLS Metathesaurus contains over 1.5 million English terms drawn from more than sixty medical vocabularies, and organized in some 800,000 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 from the semantic network.

The Gene Ontology database

The Gene Ontology™ project[4] "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 these annotation files, providing a link between gene and gene products on the one hand and the three subdomains of GO. Another feature of the database is to provide pointers to the biomedical literature when journal articles are used as evidence for the annotations.

Mapping text to knowledge bases

In the early stages of the UMLS project, techniques were developed to map text to concepts in the UMLS, resulting from the study of lexical variation in medical vocabularies. For example, methods were developed for normalizing medical terms. Normalization 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. More sophisticated techniques (e.g., the MetaMap program) were developed, allowing for partial matches. These techniques are used by various UMLS-based applications, including SemNav.

[1] http://umlsks.nlm.nih.gov ► Resources ► Semantic Navigator (free UMLS registration required)
[2] http://etbsun2.nlm.nih.gov:8000/perl/gennav.pl
[3] http://umlsks.nlm.nih.gov (free registration required)
[4] http://www.geneontology.org/

Proceedings of the ISMB'2002 SIG meeting "Bio-ontologies" 2002:29-32.

While the normalization techniques developed for medical terms can be applied to other terminologies, the characteristics of gene names may limit their usefulness in molecular biology. On one hand, these techniques may suggest similarity between terms that are not synonymous. In fact, while word order can generally be ignored in clinical terms, it often matters in molecular biology terms. For example, the terms *3'-5' exonuclease* and *5'-3' exonuclease* should not have the same normalized form. On the other hand, these techniques may fail to identify similarity among gene names or symbols. For example, although naming homologous gene products, the two symbols *FGF1_HUMAN* and *Fgf1* do not result in the same normalized form. Neither do the corresponding names *Heparin-binding growth factor 1 precursor* and *Fibroblast growth factor 1 (heparin binding)*. In GenNav, so far, we use simple but limited string matching techniques (exact or partial) on gene names and symbols, including the synonyms.

Visualization

The visualization of complex and extensive knowledge structures poses specific challenges, namely to represent enough significant information while limiting the cognitive load. While powerful methods for browsing the literature have been developed by the information retrieval (IR) community and are now implemented in virtually any IR system, little attention has been paid to visualization techniques developed for displaying complex knowledge structures. For example, the tree metaphor, well adapted to the representation of single inheritance hierarchical structures such as file systems, is also often used for representing polyhierarchical structures and directed acyclic graphs, imposing a significant cognitive effort on users for recreating a graph from multiple trees. In both SemNav and GenNav, we use the graph visualization package GraphViz[5] to generate graphical representations of multiple inheritance. However, while we take advantage of graphical representation for displaying hierarchies, we still use lists to represent concepts in associative relationship (e.g., concepts co-occurring with a given disease, and gene products annotated with a given molecular function). Concepts in associative relationship are often many, and the justification for their spatial location on the graph is less obvious.

Another element for limiting the cognitive load is to reduce the complexity of what needs to be represented. In SemNav, a transitive reduction is performed on complex graphs in order to limit the hierarchical relationships displayed to those that cannot be inferred by transitivity. It is also often possible to restrict the concept space to a smaller space by focusing on certain characteristics, e.g., a given vocabulary in SemNav or a given species in GenNav. Finally, the features to be displayed may also be selected by users.

Navigation

Navigation features allow users to adopt an exploratory attitude while getting involved with the knowledge. In particular, by navigating, users can explore the associations represented in the knowledge base (e.g., between diseases and manifestations, between gene products and GO concepts). To be fully powerful, the navigation must result in sliding a window on the data, i.e., in providing the user with a different perspective on the data. For example, in GenNav, starting from the space of gene products, a user is first presented with the GO concepts annotating this gene product. By selecting any of these GO concepts, the user is transported into the space of GO concepts, from which all gene products annotated with the GO concept of interest are displayed. In this exploratory mode, users can formulate hypotheses to be tested by more exploration or experiments.

Navigation may also lead to external databases. For example, the GO database contains pointers to Pubmed, an interface to the MEDLINE® bibliographic database. These links can be activated in GenNav and other GO browsers, allowing users to access the documents supporting the annotations of gene products with GO concepts.

Most of the lessons learned while developing SemNav (for browsing general biomedical knowledge) were applicable to GenNav (for browsing molecular biology knowledge). However, the lexical techniques suitable for mapping text to clinical terminologies require adaptation to the specificity of molecular biology terminologies.

[5] http://www.graphviz.org/