A Semantic Network for Modeling Biological Knowledge in Multiple Databases

Jeffrey Stone
University of Vermont

Xindong Wu
University of Vermont

Marc Greenblatt
University of Vermont

Follow this and additional works at: https://scholarworks.lib.csusb.edu/ciima

Part of the Management Information Systems Commons

Recommended Citation
Stone, Jeffrey; Wu, Xindong; and Greenblatt, Marc (2004) "A Semantic Network for Modeling Biological Knowledge in Multiple Databases," Communications of the IIMA: Vol. 4 : Iss. 4 , Article 5.
Available at: https://scholarworks.lib.csusb.edu/ciima/vol4/iss4/5

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Communications of the IIMA by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
A Semantic Network for Modeling Biological Knowledge in Multiple Databases

Jeffrey Stone
Department of Computer Science, University of Vermont, Burlington, VT 05405
(802) 238-8201 Fax: (802) 656-0696, jstone@emba.uvm.edu

Xindong Wu
Department of Computer Science, University of Vermont, Burlington, VT 05405
(802) 656-7839 Fax: (802) 656-0696, xwu@cs.uvm.edu

Marc Greenblatt
Department of Medicine, University of Vermont, Burlington, VT 05405
(802) 656-7839 Fax: (802) 656-0696, Marc.GreenBlatt@vtmednet.org

ABSTRACT

We have developed a semantic network of biological terminology to aid in the retrieval and integration of biological information from a variety of disparate information sources. Our semantic network strives to provide a categorization of biological concepts and relationships among these concepts. The semantic network will impart a knowledge structure through which computers can “reason” and draw conclusions about biological data objects and will provide a federated view of the many disparate databases of interest to biologists. In the development of our system, we have included the concepts from several established controlled vocabularies, chief among them being the National Library of Medicine’s Unified Medical language System (UMLS). While the UMLS Metathesaurus provides an excellent controlled vocabulary, we have found their semantic network lacking in sufficient detail to be useful as a tool for categorization of biological concepts in databases. We would like to provide a categorization of concepts that provides finer detail than their semantic network without the considerable size and complexity of their Metathesaurus. Our complete semantic network consists of 183 semantic types and 69 relationships.
Recent advances in the fields of computational biology, cloning and genetics have led to vast amounts of data, which are providing an unprecedented volume of knowledge to researchers and medical personnel. This information will be critical for the understanding of biological processes and structure and has allowed for the development of new treatment approaches for disease such as gene therapy and pharmacogenetics. However, the amount of data that the average researcher must comb through on a daily basis has become unmanageable.

Although advances in processing power, data storage, and search algorithms have resulted in great improvements in traditional search engines, sites such as Google and Yahoo are still of little use to the biologist searching for domain specific information. This is because traditional search engines rely on keyword processing alone without context. By utilizing the semantics of the words instead of the keywords themselves we can greatly enhance the search process.

There has been much interest in the idea of a ‘Semantic Web’ as defined by Tim Berners-Lee in 1998 (Berners-Lee, 1998) (Fensal, 2003). Although a general semantic system encompassing all information available on the Internet would ideal, that day is still far in the future. We can however build smaller, domain specific semantic systems for use today. We have designed a semantic network (Griffith, 1982) (Feng et al., 2002) of biological information which software developers can utilize to better serve the researcher in managing this vast amount of data.

In this paper we will describe the National Library of Medicine’s Unified Medical Language System (UMLS) Metathesaurus and their semantic network and their respective strengths and shortcomings. We will then describe a hybrid system based on their semantic network augmented with some of the ontology of the Gene Ontology Consortium’s controlled vocabulary, which has been incorporated into the UMLS Metathesaurus, and several new concepts and relationships of our own.

Starting with the UMLS semantic network, we have made additions which have resulted in 65 new semantic types and 15 new relationships. Many of these new semantic types are from the Gene Ontology Consortium’s controlled vocabulary and are used to classify genomic data. To decide on what items to include in our semantic network we used the approach of looking at the common types of databases used by biological researchers to decide which semantic types and relationships could be added. We also found that several of the concepts in the original UMLS semantic network were redundant or not of importance to our target databases. Through this approach we were able to eliminate 16 semantic types without adversely affecting the network. We then show how the items in disparate database systems will fit into our new semantic network. We will conclude with a brief discussion on how our semantic network will be used to classify items and user interests in an intelligent digital library system.
RELATED WORK

Although there have been several ontologies developed for describing biological data, there is still no published knowledge base that can be used to cover the number of disparate databases which are used by biomedical professionals. Yu et al (1999) adapted the UMLS semantic network to cover genomic knowledge and Hafner et al (1994) also used the UMLS as a basic building block for their system of representing biomedical literature. Most other biomedical resource systems such as Genbank and the Protein Data Bank (PDB) contain crucial facts, but do not contain information about the concepts and relationships of the many inter-related terms (PDB).

The Gene Ontology Consortium has developed a large controlled vocabulary for the unification of a genetic concepts and terminology. This controlled vocabulary along with several others is now part of the massive UMLS Metathesaurus. These ontologies provide the vocabulary for the description of many biological concepts such as the annotation of the molecular function, biological process, and cellular component of gene products. This metathesaurus is a big step towards the unification of biological knowledge, however, it is simply far too complex to provide a federated solution to unifying biological databases.

The structure of the Gene Ontology vocabulary provides a good example of the vocabularies that make up the UMLS Metathesaurus. The Gene Ontology controlled vocabulary is based on the annotation of gene products. A gene product is a physical entity. Gene products may be RNA or proteins. These gene products may have many molecular functions. A molecular function is a description of what a gene product does. One drawback of the Gene Ontology system is that the molecular function only describes what a gene product has the potential to do without regard to where or when this function may take place. Such semantics as to where and when a function takes place could be contained within a semantic network.

The National Library of Medicine has a long-term project to build a Unified Medical Language System (UMLS) that is comprised of three major parts: the UMLS Metathesaurus, SPECIALIST Lexicon, and the UMLS Semantic Network. The Metathesaurus provides a large integrated distribution of over 100 biomedical vocabularies and classifications. The Lexicon contains syntactic information for many terms, component words and English words, including verbs, not contained in the Metathesaurus. The Semantic Network contains information about the types or categories to which all Metathesaurus concepts have been assigned and the permissible relationships among these types (UMLS). The UMLS system has been used successfully in many applications mostly involving scientific literature.

The UMLS Semantic Network provides an ideal framework for federating disparate databases. However, the current structure of the UMLS Semantic Network is most useful for scientific literature and clinical trial information. If one is trying to use the UMLS Semantic Network for federation of several disparate databases, they will find the network is not sufficiently broad to cover the multiple items in all of these databases.
We have therefore decided that to best suit the needs of our digital library system, we must develop our own controlled language system. To do this, we have started with the basic framework of the UMLS semantic network and then pruned some of the less important details and added new concepts and relationships where needed to cover the databases in our digital library.

**SEMANTIC NETWORK STRUCTURE**

Our semantic network will be comprised of nodes representing semantic types and relationships between these nodes. Each node represents a category of either a biological entity or an event. The entities and events used in our semantic network result from a merging of some of the concept names in the National Library of Medicine’s Unified Medical Language System and the Gene Ontology Consortium’s controlled vocabulary.

Most relationships in our system will be of the is-a variety, such as a human is-a organism. However, many biological entities do not fit into a simple hierarchical structure. Therefore we need additional relationships between multiple hierarchies to accurately represent the complexity of biological data. These interconnecting relationships and hierarchies make up our semantic network.

The first major entity category is that of an organism. This represents a simple taxonomic hierarchy of organisms. Another category is that of anatomical structure. This hierarchy represents embryonic structures, anatomical abnormalities, body parts, organs, organ components, tissues, cells and cellular components including genes. The cellular component hierarchy will be mostly taken from the Gene Ontology Consortium’s hierarchy. A third major category is that of a conceptual entity. This category will include items such as temporal, qualitative, quantitative, functional and spatial concepts. There is also a category for medical findings including symptoms and laboratory results.

In addition to entities, we also have categories of events including activities, phenomenon and processes. Activities include such things as health care activities such as laboratory, diagnostic, therapeutic and preventative procedures, and research activities, such as research techniques and methods. The Phenomenon or Process category includes biological functions and pathologic functions. Biological functions include physiologic functions such as organ or tissue functions, cellular functions or sub-cellular component function and molecular functions such as genetic function.

The events category is a crucial component of our semantic network since the information in many of the most important databases of interests to biologist relate to the information in this category. This is also the most difficult category to design due to the lack of a clear hierarchical structure to events. Again, we have borrowed from the Gene Ontology Consortium to develop the molecular and biological functions, however, we have chosen to truncate the tree structure of their system to prevent the relationships between these functions from getting too complex.
The relationships that tie all of these hierarchies together complete our semantic network. These relationship links between the hierarchies allow us to represent knowledge about an entity or an event. For example we may represent a gene as a cellular component that is in the hierarchy of anatomical structures. This gene will produce a gene product. That gene product is also a cellular component that may have a biological function and possibly a molecular function. The gene may be part of many different organisms and it may be associated with a pathological function.

Initially we are starting with very basic relationships among these hierarchies. We will rely on only top-level relationships such as the is-a relationships that make up the various hierarchies and the associated-with relationships that tie these hierarchies together. We will also build the next layer of relationships below the associated-with layer. This will comprise of physically-related-to, spatially-related-to, functionally-related-to, temporally-related-to and conceptually-related-to relationships. These relationship links have been built through a restructuring of the UMLS concepts and the Gene Ontology Consortium’s hierarchy.

Our semantic network is similar in structure to the UMLS system, but is able to classify the biological information in far greater detail. This is especially true with genomic data. The UMLS system was designed by the National Library of Medicine and has naturally taken the view of that institution on how to classify data. We have focused more on the end users and how they would view the data. Therefore we have removed many of the nodes that have to deal with government regulation, legal information and health care institution information and have focused more on pure biomedical research information. Other controlled vocabularies are specific for one branch of biomedical research such as the genomic research modeled by the Gene Ontology Consortium. Our system is based not on the research areas themselves, but rather the data that will be included in our digital library system. Therefore, our system will evolve over time as more items are added to our digital library.

**Dictionary terms reside at each node**

Every node in our system will have a list of distinct concept classes. Each distinct concept class will have a list of synonymous words and phrases. These terms are primarily obtained from the Medical Subject Headings (MeSH) compiled by the National Library of Medicine (NLM). Every separate meaning will appear as it’s own concept class, but a node may have multiple concept classes. All of concept classes taken together will contain the entire set of terms in our dictionary. It is at this level that each item in our digital library will be classified into our semantic network.

Every entry in our digital library will have a list of these terms associated with it. Most items in biological databases are designed for keyword-based queries and therefore already have this information associated with them. In the future, the possibility exists for extracting this information from text sources as well (Craven & Kumlien, 1999).
Decisions on what concepts and relationships to include

As stated earlier, we have started with the basic structure of the UMLS system. Starting with this system we remove those items that are too detailed to be included in such a system by manually pruning the “Entity” and “Associated-with” hierarchies. This careful pruning is done with a base set of databases in mind. These include the popular Protein Data Bank (PDB), the Online Mendelian Inheritance in Man (OMIM) and mutation databases for the p53 and CDKN2a (p16) tumor suppressor genes (OMIM)(p53DB)(CDKN2a) to demonstrate our networks usefulness with private data.

Using the databases, we now identified the corresponding types in our truncated UMLS semantic network along with any concepts not included by manual inspection. Where no concepts are included, we added new types and determined where they should be placed in the semantic network (fig. 1 and fig. 2).

We have found that many of the Entity Semantic Types of the UMLS semantic network are beyond the scope of our project. We have therefore performed a careful manual pruning of the network to remove those nodes that are not of interest. Most of the items removed pertained to specific medical equipment and physical health care facilities. We removed the node for manufactured object and all children of this node. However, since the node for clinical drug fell under this node, we would have to re-insert this node elsewhere in the network. The most logical place for this is a new node under chemical substance. We also removed the nodes of Finding, and several of the sub-nodes under the Event category such as a machine activity, and educational activity.

Figure 1: Shown here is a simplified hierarchy showing a portion of our “entity” semantic types. Each node represents a category of biological concepts. At each node will reside one or more concept classes, which will contain different terminology with the same or similar meaning. The hierarchical structure is represented by means of “is-a” linkages. The rectangular boxes come from the National Library of Medicine’s UMLS project. Oval nodes are new types that come from different ontologies outside of the UMLS project as well as types that we have designed ourselves.
Figure 2: Another important semantic type is that of an “event”. Many of the added nodes for the “event” type originate from the Gene Ontology Consortium’s controlled vocabulary. The hierarchy shown is only a small portion of the entire event hierarchy. Each child of the “event” type has several children, many of which have several children of their own.

Figure 3: Besides the “is-a” relationships that represent a hierarchical structure, we also have “associate_with” relationships that can represent the many non-hierarchical relationships that biological items may have to one another. The importance of these relationships is one of the reasons why we chose a semantic network to represent the terms in our dictionary.
We inspected likewise the semantic relationships of the UMLS system for areas to prune. We found less to prune here, but there were a few items, such as evaluation-of, analyzes, assesses-effect-of, and measures.

The information contained in the Protein Data Bank is primarily structural data of proteins. However, the current UMLS semantic network does not contain structural information. We therefore have added a node for Protein Structure under the Anatomical Structure Node. This new node will have 4 child nodes for primary, secondary, tertiary and quaternary structure protein structures. The typical item in the PDB will be a “3-D Structure” and it will have an associated “1-D Structure” and a “2-D Structure”. Items within the PDB might also have the relationship of being similar to another protein’s structure or function. We therefore added semantic relationships for similarly-related-to, with its child nodes of physically-similar-to and functionally-similar-to.

We used a similar approach with information contained in OMIM- Online Mendelian Inheritance in Man [OMIM]. This database of genetic disorders in man is rich in information, however, most of this information is structured in the form of text documents. This creates some difficulty in mapping the information to the semantic network. Nonetheless, there is some basic information available on each document that can be searched efficiently. This information includes allelic variants, gene map disorders, and clinical synopsis, and references. The allelic variants are a physical relationship to whereas the clinical synopsis fits into the causes relationship and also under the disease or syndrome event. Much of the information within the OMIM database would fit nicely under the Gene Ontology Consortiums controlled vocabulary, which has been incorporated into our system.

USES OF OUR SEMANTIC NETWORK: A DICTIONARY FOR A DIGITAL LIBRARY

To aid researchers in obtaining, organizing and managing biological data, we have proposed a sophisticated digital library system that utilizes advanced data mining techniques. Our digital library system will be centralized with Web links to data repositories physically located on the Web. Our digital library will be based on a framework used for conventional libraries and an object-oriented paradigm, and will provide personalized user-centered services based on the user’s areas of interests and preferences.

This approach begins from the centralized, structured view of a conventional library, and seeks to provide access to the digital library through electronic means including the Web, while maintaining the advantages of decentralization, rapid evolution and flexibility of the Web. The core of our project will be the knowledge object modeling of data repositories, and an agent architecture that provides advanced services by combining data mining capabilities.

The knowledge objects are defined to be an integration of the object-oriented paradigm with rules. The proper integration provides a flexible and powerful environment, as rule-based components provide facilities for deductive retrieval and pattern matching, and object-oriented components provide a clear intuitive structure for programs in the form of class hierarchies. The
design criteria of the model will be completeness, compactness, and simplicity. It will allow the mapping of all types of biological data. The classes will account for any type of biological items and their relationships among them. Each offspring class will be a merge of many detailed parts that are to be composed in the form of a URL list to describe the biological information under consideration. The model will thus be both complete and compact, since it covers all biological data within the scope while the number of classes is kept minimal. This will provide a solid base, making the model robust to changes and simple to use.

figure 4: An overall schema of our digital library system. The HCV induction engine will be the “brain” of the discovery agent. It will take two input sets of documents; one set the user has seen, and the other the user has not visited. It will generate rules in the form of conjunctions of keywords in the dictionary to identify the user’s areas of interest, and forward the rules to the user profiles. The Dictionary component of our digital library will be provided by the terminology contained at the nodes of our semantic network.

To make personalized service possible, a “user profile” that represents the preferences of an individual user can be constructed based upon the users past activities, goals indicated by the
user, and options. Utilizing these user profiles, our system will make relevant information available to the user in an appropriate form, amount, and level of detail, and especially with minimal user effort.

One crucial component of our digital library system will be a dictionary of biological terminology. This dictionary will play an important role in building up user profiles. Advanced knowledge discovery agents can then utilize these user profiles to learn about a user's area of interests and to guide the user in searches of the databases. This dictionary will also be used in the development of categorization rules for each biological object in our digital library as well as for indexing the database.

In the construction of the dictionary, we are presented with some difficulties due to the nature of biological data. Some of the problems encountered are multiple names for the same protein or gene in different organisms, the dependency of the biological state in which the function is taking place and multiple functions for the same protein. These problems preclude the use of a simple hierarchical dictionary structure. However, by constructing our dictionary as a semantic network utilizing a directed graph based paradigm, we can overcome these obstacles and provide a model that can accurately model the information contained in multiple biological databases.

Figure 5: Shown here is a partial schema of the overall semantic network. Solid lines are "is-a" links whereas the dashed lines indicate a category of "associate-with" relationships. A user friendly interface is being developed through which the user may browse the semantic network or enter terms to find relationships to these terms.
COMPARISONS

The current UMLS Semantic Network has 134 semantic types and 54 relationships to link the semantic network together. Using the UMLS network as a starting point, we removed 16 semantic types that we found either ambiguous, redundant or outside the scope of our project based on our target audience of biomedical researchers. To this network we added 65 semantic types, 55 of which came from the Gene Ontology Consortium’s controlled vocabulary. Many of these new types are either molecular functions or biological processes. We have also added 15 new relationships. The semantic types and relationships are listed in their entirety in appendices 1 and 2. Although the complete semantic network is too complicated to be shown in its entirety, a simplified schema is shown in figure 5.

CONCLUSIONS

We believe that by restructuring the UMLS semantic network and adding to it as needed, we can create a new semantic network that can effectively cover the many disparate biological databases that one would want to include in a digital library system. As we add more databases to our digital library, our semantic network will grow over time. However, by carefully inspecting the nodes, we should be able to manage that growth and ensure that we maintain a balance between covering all the data in our system and avoiding the fine details that will become useless in the larger system.

APENDIX 1. SEMANTIC TYPES

Entity
  Physical Object
    Organism
      Plant
        Alga
        Fungus
        Virus
        Rickettsia or Chlamydia
        Bacterium
        Archaeon
        Animal
          Invertebrate
            Vertebrate
              Amphibian
              Bird
              Fish
              Reptile
              Mammal
                Human
                  Mouse
Anatomical Structure

Embryonic Structure

Protein Structure

Primary Structure
Secondary Structure
Tertiary Structure
Quartenary Structure

Anatomical Abnormality

Congenital Abnormality
Acquired Abnormality

Fully Formed Anatomical Structure

Body Part, Organ, or Organ Component
Tissue
Cell

Cellular Component

Cell Type
Extracellular
Unknown

Gene or Genome

Substance

Chemical

Chemical Viewed Functionally
Pharmacologic Substance
Antibiotic
Clinical Drug
Biomedical Material
Biologically Active Substance
Neuroactive Substance or Biogenic Amine
Hormone
Enzyme
Vitamin
Immunologic Factor
Receptor
Indicator, Reagent, or Diagnostic Aid
Hazardous or Poisonous Substance

Chemical Viewed Structurally
Complex
Organic Chemical
Nucleic Acid, Nucleoside, or Nucleotide
Organophosphorus Compound
Amino Acid, Peptide, or Protein
Carbohydrate
Lipid
Steroid
Eicosanoid
Inorganic Chemical
Element, Ion, or Isotope

Body Substance
Food
[Entity] (continued)

Conceptual Entity
  Idea or Concept
    Temporal Concept
    Qualitative Concept
    Quantitative Concept
    Functional Concept
  Body System
    Biochemical Cascade or Cycle
  Spatial Concept
    Body Space or Junction
    Body Location or Region
    Molecular Sequence
      Nucleotide Sequence
      Amino Acid Sequence
      Carbohydrate Sequence

Finding
  Laboratory or Test Result
  Sign or Symptom

Organism Attribute
  Clinical Attribute

Organization
  Health Care Related Organization
  Professional Society
  Self help or Relief Organization

Group
  Population Group
  Family Group
  Age Group
  Patient or Disabled Group

Event
  Activity
    Behavior
      Social Behavior
      Individual Behavior
    Daily or Recreational Activity

  Occupational Activity
    Health Care Activity
      Laboratory Procedure
      Diagnostic Procedure
      Therapeutic or Preventive Procedure

  Research Activity
    Molecular Biology Research Technique
    Biological Procedure
    Chemical Procedure

Phenomenon or Process
  Human caused Phenomenon or Process
  Environmental Effect of Humans
  Natural Phenomenon or Process
    Biological Function
      Behavior
[Event](continued)

[Phenomenon or Process](continued)

[Natural Phenomenon or Process](continued)

[Biological Function](continued)

Cellular Process
- Cell Communication
- Cell Death
- Cell Differentiation
- Cell Growth and/or Maintenance
- Cell Motility
- Membrane Fusion

Development
Physiological Process
Viral Life Cycle
Organ or Tissue Function

Molecular Function
- Anticoagulant activity
- Antifreeze activity
- Antioxidant Activity
- Apoptosis Regulator Activity

Binding
- Amino Acid Binding
- Antigen Binding
- Carbohydrate Binding
- Cofactor Binding
- Drug Binding
- Glycosaminoglycan Binding
- Hormone Binding
- Host Cell Surface Binding
- Isoprenoid Binding
- Lipid Binding

[Binding](continued)
- Lipopolysaccharide Binding
- Metal Ion Binding
- Neurotransmitter Binding
- Nucleotide Binding
- Oxygen Binding
- Peptide Binding
- Protein Binding
- Receptor Binding
- Steroid Binding
- Vitamin Binding

Catalytic Activity
Cell Adhesion Molecule Activity
Chaperone Activity
Immune Activity
Enzyme Regulator Activity
Motor Activity
Protein Stabilization Activity
Signal Transducer Activity

Structural Molecule Activity
- Toxin Activity
- Transcription Regulatory Activity
- Translation Regulatory Activity
APENDIX 2. SEMANTIC RELATIONSHIPS

Is a
Associated with
Physically related to
Part of
Consists of
Contains
Connected to
Interconnects
Branch of
Tributary of
Ingredient of
Spatially related to
Location of
Adjacent to
Surrounds
Transverses
Functionally related to
Affects
Manages
Treats
Disrupts
Complicates
Interacts with
Prevents
Activates
Promotes
Deactivates
Brings about
Produces
Causes
Create Bond
Break Bond
Releases
Signals
Transports
REFERENCES


Fensal et al. (2003) Spinning the Semantic Web: bringing the world wide web to its full potential. MIT Press, Cambridge, MA.


PDB. The Protein Data Bank. http://www.pdb.org


