AN INTELLIGENT DIGITAL LIBRARY SYSTEM FOR BIOLOGICAL DATA

A Thesis Presented

by

Jeffrey E. Stone

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Computer Science

February, 2005

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Master of Science, specializing in Computer Science.

Thesis examination Committee:

________________________Advisor
Xindong Wu, Ph.D.

________________________ Co-Advisor
Marc Greenblatt, M.D.

________________________ Chairperson
Jim Vigoreaux, Ph.D.

________________________ Vice President for
Frances E. Carr, Ph.D. Research and Dean
Of the Graduate College

Date: October 14, 2004

Abstract

To aid researchers in obtaining, organizing and managing biological data, we have developed a sophisticated digital library system that utilizes advanced data mining techniques [Stone et al 2004a]. Our digital library system is implemented as a centralized J2EE web application with links to publicly accessible data repositories on the Internet. The digital library is based on a framework used for conventional libraries and an object-oriented paradigm, and provides personalized user-centered services based on the user's areas of interests and preferences. To make personalized service possible, a "user profile" that represents the preferences of an individual user is constructed based upon a user's past activities, goals indicated by the user, and options. Utilizing these user profiles, our system makes relevant information available to the user in an appropriate form, amount, and level of detail with minimal user effort. The core of our project is an agent architecture that provides advanced services by combining data mining capabilities with domain knowledge in the form of a semantic network [Stone et al 2004b]. The semantic network imparts a knowledge structure through which the system can "reason" and draw conclusions about biological data objects and provides a federated view of the many disparate databases of interest to biologists. In the development of our semantic network, 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). Our complete semantic network consists of 183 semantic types and 69 relationships.

Acknowlegements

I would like to thank my advisor, Dr. Xindong Wu, for introducing me to the field of data mining and his insight on novel applications of this technology. I also thank my co-advisor, Dr. Marc Greenblatt for his expert medical and database advice. Finally I would like to thank Dr. Jim Vigoreaux for his help in completing my final defense.

This study is supported by a DOE EPSCoR grant.

Table of Contents

Table of Figures

Chapter 1 Introduction

1.1 Motivation

The burgeoning amount of biological information (e.g., the published literature, genome and proteome projects) is confronting researchers with challenges in dealing with this large volume of complicated data. With the advance of technologies in both the biological and computer sciences, universities and research centers are producing a huge amount of experimental data, diverse new discoveries, and related publications. For instance, the BioNetbook [BioNetbook] has recognized and collected biological information databases around the world, now totaling 1750 and growing. In addition, they have recognized 8048 relevant web pages including bibliographical, analysis tools, software and courses in biological research. In their 2004 annual database issue, Nucleic Acids Research has recognized biological information storage sites around the world, classified them into 11 categories of biological information, and listed 548 web sites under operation [Galperin 2004]. These databases, dispersed in different locations, serve important roles in modern research including the storage of experimental data, the maintenance of information, and the integration of diverse data sources. Researchers often review multiple publications and other databases to arrive at a comprehensive understanding and generate or validate their hypotheses. In this model, biological phenomena may be viewed as being composed of a number of sub-disciplines (e.g., structural biology, genomics, proteomics, biochemistry). However, to obtain a coherent picture of biological phenomena at the molecular, cellular, and organism levels, one must look both at all of these attributes and the relationships among them. To do that currently requires finding which databases contain the relevant information and then searching through the databases one by one.

Biological information is a knowledge intensive subject. Some web sites provide biological information by listing a number of databases, arranged by category, and entrust the user with all the search responsibilities. Although public data repositories such as the National Center for Biotechnology Information (NCBI) and the Protein Data Bank (PDB) integrate several databases on one site and contain much of the publicly held biological data, these sites still face difficult problems caused by a flood of new and diverse data and variation in format. In addition, they suffer from a lack of robustness in search techniques for highly interconnected databases, such as being stranded during a series of search processes. Some requests inherently require going through a complicated serial search on highly inter-related databases. Under these circumstances, researchers have a hard time locating useful information in an efficient way and keeping informed of updates. As data acquisition from new experimental techniques (such as micro-arrays) flourishes, the problem of finding the right information gets worse.

Sometimes researchers make use of modeling techniques and domain knowledge to fully utilize the known information in solving complex problems. Knowledge intensive applications, such as biological information services, would be good candidates for modeling knowledge. With the models or domain knowledge the system could greatly improve its retrieval performance. These modeling techniques have attracted researchers from many practical fields and have been exploited for reasoning about problem solving because they can explicitly represent the underlying knowledge of the composition of the target domain and how components of the model are related and interacted upon. This explicit representation of underlying knowledge would bridge the possible gap between two characteristically different disciplines: biological information processing and computer software development.

Unlike general search engines used on the Internet, biological information searching is a specific application domain. Although search engines such as Google may return many positive results, these pages are ranked by popularity, not scientific merit. Since only a limited number of researchers are interested in this application, it is possible that each individual can be better served by an interface customized to their particular interest. To this end, agent technology that has become very popular in other web applications can be benevolent to users. In this section, we discuss object modeling, user profiles, and agent technology. The complexity and diversity of information can be approached by object modeling and user profiles with agent technology can be used for user centered services.

1.2 How Digital Libraries Can Help Manage Data

The days of the card catalogue in the local library are long gone. Even the smallest of town libraries have computerized search engines to find and organize the wealth of information on their shelves. The birth and evolution of the World Wide Web, as haphazard and uncontrolled as it is, have allowed users in rural America to access many of the same resources that had previously only been available in large metropolitan areas.

As advantageous as the “Information Age” has been, managing this ever increasing body of information has proved to be quite difficult. While the conventional “brick and mortar” libraries will continue to play an important role in our society, the need for better means to control information will be among the most important and influential pursuits of this century. Digital libraries, both small and large, promise to play an important role in the pursuit of information management.

A digital library is, like a traditional library, a collection of books and reference materials. Unlike a traditional library, however, the collection of a digital library is, as you would expect, digital, and is usually served over the World Wide Web. The Working Group on Digital Library Metrics further characterizes digital libraries as providing the collection of services and the collection of information objects that support users in dealing with information objects and the organization and presentation of those objects available directly or indirectly via electronic/digital means [D-Lib].

We believe that most digital libraries currently available are missing one of the most important features of traditional libraries: the librarian. If we know what we want from the library, we can walk in, look it up in their computer search engine and walk right to it on the shelves. However, where traditional libraries really shine is when you are not really sure what you are looking for. The librarian is trained to help people who have a vague idea of what they want, but don't remember the name, don't know what is in the library, or simply don't know enough about the subject to productively search using standard search engines.

Intelligent agents that combine user profiling, data mining, and artificial intelligence techniques can be developed to provide many of the services that have traditionally been offered by the librarian. This is the area of focus where our digital library is different from any others [Stone et al 2004a]. We use these advanced data-mining techniques to guide the researcher in their navigation of the library. The system will learn about each user and then be able to recommend items or suggest alternative terms to search under.

1.3 Summary of Results

Our approach began with a centralized, structured view of a conventional library, and provides access to the digital library via the Internet, thus maintaining the advantages of decentralization, rapid evolution and flexibility of the Web. The application is a J2EE web application built using the Model View Controller (MVC) architecture on a Tomcat server with JDBC connections to a MySQL database. The core of our project is a 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 of which provides a flexible and powerful environment for deductive retrieval and pattern matching.

To make personalized service possible, a “user profile” representing the preferences of an individual user is constructed based upon 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 is a dictionary of biological terminology. This dictionary will play an important role in building the user profiles as well as the categorization rules of each item in our digital library. 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.

To overcome these obstacles and provide a model that can accurately model the information contained in multiple biological databases, we have developed our dictionary as a semantic network of biological terminology utilizing a directed graph based paradigm [Stone et al 2004b]. Our semantic network strives to provide a categorization of biological concepts and relationships among these concepts. The semantic network imparts a knowledge structure through which our system can “reason” and draw conclusions about biological data objects. The Unified Medical language System

(UMLS) contains a large semantic network of its own that we have used as a base for our system [UMLS]. However the UMLS is in some aspects not general enough for use in categorizing multiple biomedical databases and also contains too many terms that are outside of the scope of our project. Therefore, we have trimmed some of the detail from the UMLS system and added new types and relationships to this system to provide a more general coverage of biological databases. Our complete semantic network consists of 183 semantic types and 69 relationships.

Our semantic network is 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 [GO].

The rules that are used for the recommendation of items in our system are generated with the popular open-source Weka data mining package [Witten 2000]. Regeneration of rules will take place upon the end of each session, or optionally during a session when prompted by the user. At the time of regeneration of the rules, relevant data is extracted from the user's profile and passed to the Weka J48 program. J48 uses a decision tree algorithm to generate the classification rules that will be the basis for the recommendation of items in the library. After the rules are generated, they are saved into the user's profile and used in subsequent searches to recommend items in the library.

1.3 Thesis Structure

In this chapter, I have described the problem of information overload. I then described what a digital library is and how it can help manage the enormous amount of data that we are confronted with. Chapter 2 describes the concept of object modeling and how it relates to our system. Next, I define what a semantic network is and how it can be used in a system to provide a knowledge base for our system and to control the vocabulary of our system. I will then describe our semantic network in detail. Chapter 4 begins with a discussion of common user profiling techniques and then describes the decision tree algorithm for generating user rules. I then describe the system architecture for our digital library. Chapter 6 provides a comparison of related systems and I conclude with a description of how our digital library works and some areas where it can be improved upon.

Chapter 2 Knowledge Object Modeling

Biological information can be broken down and represented in forms of objects with the integration of logic rules. Each object can have a set of rules that govern its behavior and appearance, as well as communication to other objects. Association links between objects can be represented in the form of rules in knowledge objects and used to conduct heuristic search over the databases.

The first task in developing a knowledge object model for web-based biological information is to include a means for working with different forms of media. This model will be the first step towards building a manageable system for our digital library system in which biological data can be easily stored, extended, reordered, assembled and disassembled on a component basis. The model will also make accessible properties that might be important to the user, especially in searching or classifying biological information.

The design criteria of the model will be completeness, compactness, and simplicity. In this project the model is restricted to cover a few different media types including images, and text. The number of classes, their attributes and methods will be kept to the minimum. Generic classes will be used wherever possible, inheritance will be used, composition of different building blocks may be merged into one class, and convenience functions and attributes will ideally be kept to none.

Based on previous work on object-oriented modeling of paper-based documents [Nguyen, Wu & Sajeev 1998], Figure 1 shows a preliminary design of the object diagram of our biological object model, using the Object Modeling Technique (OMT) notation [Rumbaugh et al 1991].

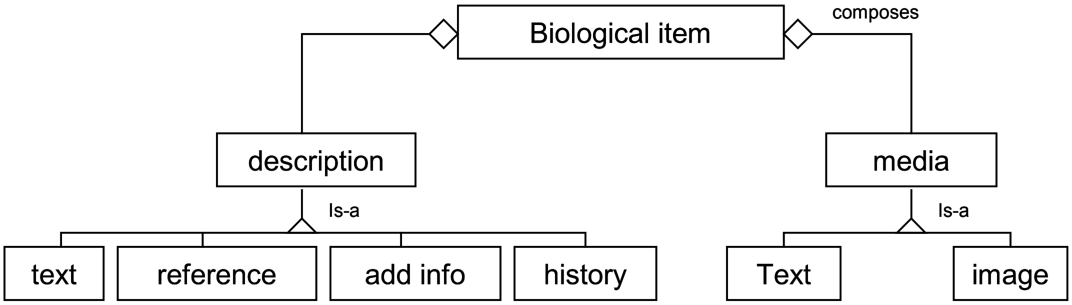


Figure 1. A schema of a knowledge object model for biological information.

According to this model, a biological item will be decomposed into knowledge objects (classes or objects for short), each holding an internal state and a well-defined behavior. The object's internal state will be defined by attributes and constraints, and its behavior by rules and methods. The biological item object will represent a variety of biological data including annotation and publications, and will be constructed by the basic building block of description. Instances of the biological item class at the top level of the hierarchy will correspond to the most general form of biological data. It will be designed to completely cover all types of biological data in digital libraries, including images and text, and ranging from page images to interactive and compound documents. The biological item object can be composed of some description objects and some media objects, or just one or the other.

A description object, inheriting the aggregation relationship to itself from the biological item, can in turn contain itself recursively. A biological item will thus be composed of a recursive chain of description objects, which can be assembled, disassembled, and reordered, allowing for the whole document to be modified, extended, or truncated on a component basis, without losing the coherence of its overall structure. Description will be the generic class for all structural components of a biological item, such as text, reference, additional information, and so on. Our model will therefore allow for organizational components such as a description containing several parts, a part containing several subparts, etc., commonly found in biological databases. Since all these structural components share the same properties and behavior, it is most appropriate to have one single representative class. This design will make the model simpler and more compact. It will give authors more flexibility when defining their own biological item since the model will not differentiate between various structural components. It will also support the interconnected nature of biological information which refers and re-uses passages and components in an integrated Web structure.

The above model will also make available presentation components, by defining them as child classes of description. A description object can thus be presented in a number of different kinds: text chunks, references, additional information, and so on.

Media objects will hold the real contents that make up the digital library. Their content can be unstructured, raw materials that will be used to fill in the biological item objects or their description components. The actual data type of the content will be defined in the offspring classes of media. It can be any multimedia type such as text or images. This content can be semantically incomplete, i.e. not meaningful to human readers. For example, it can be a fragment of text, which is a part of a paper. This semantic incompleteness will be reflected in the object design by the fact that no title or heading is required for an object of the media branch. Offspring objects of the media class, representing different media types, can therefore be inserted at arbitrary places in a biological item object. This concept will be applicable to all kinds of media types. Consider a paper with an image inserted halfway. If the paper is defined as one biological item object, then it must consist of three media objects of two broken text pieces and one image. We assume in our project that database personnel can index every document object and its parts in some way and save them in the database.

The design of our knowledge object model will allow the mapping of all types of biological data. The classes will account for any type of biological items and their relationships. Each offspring class will actually 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. New classes will be created if there are important properties or methods that need to be distinguished between them. The model will thus be both complete and compact, since it covers all biological items within the scope while the number of classes is kept to minimal. Simplicity will be achieved by compactness, and also by the fact that the model design is based on the familiar and well-developed object-oriented paradigm and technologies of database systems, with many *de facto* standards for structural and presentational components. With this solid base, the model will be robust to changes and simple to use.

A significant problem for most biological databases and libraries is that of annotation. The detailed information needed to describe biochemical processes and genetics is much more complex than the information of the taxonomy of living species that Linnaeus developed in the 18th century. We need ways to transfer this information efficiently and without propagating error. Figure 2 shows a preliminary design of the knowledge structure of our targeted biological information. This design is based on the biological structures embodied in the target databases and is represented as a class hierarchy. At the top is the most general knowledge concept that could be decomposed into detailed and related knowledge modules. Each node in this structure is represented as a biological item introduced above.

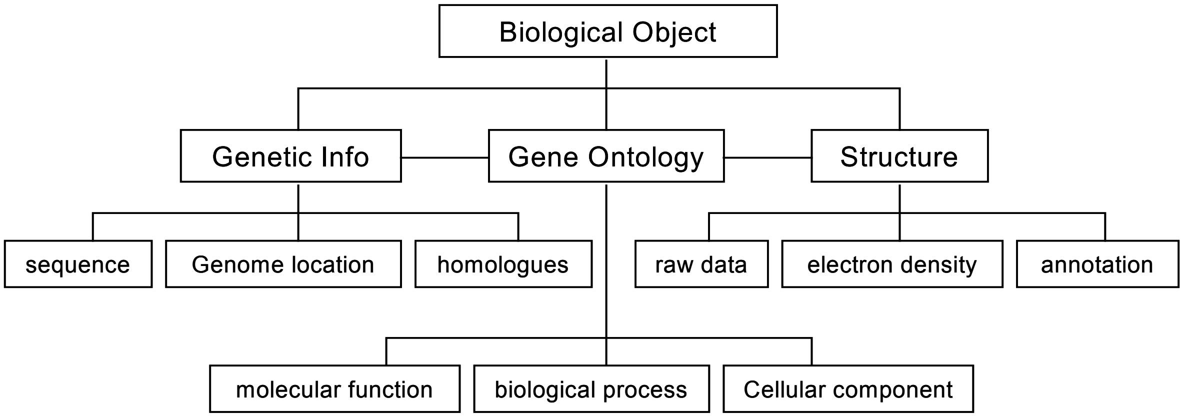


Figure 2. Biological knowledge structures.

Biologists are interested in specific genes and their products. Therefore a biological object is divided into genetic information, gene ontology, and structure information. Each of these classes is further divided into more detailed classes that give the finer details of biological information. The genetic information will contain the DNA sequence, the gene's location in the genome, and its homologues. The Gene ontology information comes from the Gene Ontology Consortium and contains information about molecular functions, biological processes and cellular components. The structure information will include raw data, electron density, and structure annotation. Other information provided will be text, references, history and information about possible relations to diseases. With this approach we will attempt to outline a biologist's view of the gene.

The design of the knowledge structure will both allow the depiction of the overall knowledge underlying the biological process and allow the system to efficiently search the right information in the databases. This structure will help a more focused, context-based retrieval over the dispersed databases.

2.1 Dealing With Documents Prepared in Different Formats.

Biological items linked in our digital library system are simple web links, but our object model approach could be further developed to allow their content to be assembled into a composite web page. The objects would be allowed to take a selection of different popular formats (such as HTML, SGML and XML), and conversion from one format to another will not be necessary. This would provide authors with flexibility in preparing their data. Allowing items in different formats requires the agents in the digital library system to be familiar with these formats. For example, the indexing agent will traverse and index all relevant items linked to the digital library system. Future projects will design an interface for each allowed format to facilitate the agents to visit and process items in these formats. Since systems exist that can convert their files to HTML, it seems possible to study our research issues using only HTML in this project. However, making everyone use HTML (or any other format) would limit the presentation medium substantially and increase the overhead for authors. The premise of the WWW and Internet was to allow heterogeneous formats. The idea of allowing different formats in this project is to free the authors of any indexing burden and make formats, like images, all indexable.

Unfortunately, not all resources are indexable. For sites that are not indexable, our library will use the entry page as the resource. While this may not seem like the ideal solution, there is in fact, a considerable need for a way to search for appropriate databases. Most people know about the NCBI site, but there are literally thousands of relevant biological portals available now. Many of these go unnoticed by the vast majority of researchers.

2.2 Deriving the Knowledge Structure From Biologists.

The knowledge structure will need to be designed through consultation between biologists and computer scientists. The need for this representation quickly increases with data complexity. This is due to many factors including the absence of permanent and unique identification of the objects, the existence of many-to-many relationships between different biological concepts such as genes and proteins and their relations to diseases and drugs, the quickly evolving knowledge of biological phenomena and data, the complex relationships between biological elements and phenomena, and the large variety of biological data types. The knowledge structure would have categories such as resource type, organism, biological domain, relationships and associations. Some nodes might contain many or even all of these categories where others may have only one or two (*figure 3*).

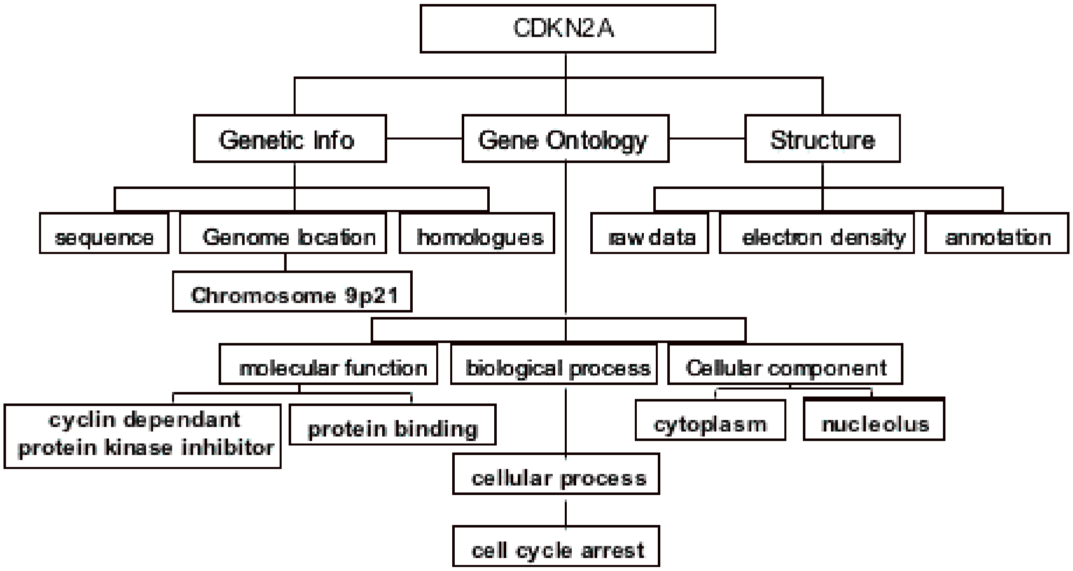


Figure 3. A knowledge structure for the CDKN2A gene.

A hierarchy structure will further define each of these subcategories. For example, when a document about a specific gene is included in our system the biological domain would be genomic. However, we can further classify a gene by molecular functions, biological processes, and cellular components. Each of these in turn can be subdivided further. Molecular functions could be any number of functions such as transcription factors, cell adhesion molecule, etc.

We will be using the Unified Medical Language System Semantic Network as a base for our knowledge structure. To this base, we will be adding additional nodes where appropriate to better describe the disparate databases used in our system. Where applicable we will use established open-source schemas such as the Gene Ontology Consortium and the Object Protocol Model (OPM).

The knowledge object modeling in [Wu & Cai 2000] is an integration of the object-oriented paradigm with logic rules, in which the class and inheritance features of the object-oriented paradigm can assist users to describe and define biological information more naturally and to emphasize the semantic rather than the syntactic content of applications. The encapsulation and dynamic binding of the object-oriented paradigm will make the resultant software more maintainable, adaptable and recyclable. Rule-based programming, meanwhile, can express constraints within and between biological information components very efficiently in its rules and provides inference power to the system. The proper integration of the object-oriented paradigm and rule-based programming will provide 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 documents in the form of class hierarchies.

2.3 Alternatives To Knowledge Objects.

There have been many research efforts reported in the literature (such as [May & Lausen 2000, Ludascher et al 1998, Bouziane & Hsu 1997] on the integration of the rule-based paradigm with object-oriented programming and on coupling rule bases and databases using either the entity-relationship or object-oriented data models, as thoroughly reviewed in [Wu & Cai 2000]. There have also been quite a number of commercial software packages (such as KEE, Prolog++ [Moss 1994], ILOG Rules [ILOG 1998] and CLIPS [Giarratano 1993]) that support both rules and objects. Among others, agent-oriented programming [Shoham 1993] and frame-based systems [Fikes & Kehler 1985] are along the line of incorporating rules into objects. However, in our knowledge object modeling, we do not specialize the object-oriented features in any way, in contrast to agent-oriented programming and frame-based systems. We will also provide efficient representation and inference mechanisms based on our existing work [Wu & Cai 2000].

In knowledge objects, rules access objects' internal states and their global contexts. The rules in knowledge objects derive new data from existing data as well as expressing constraints of the object model. Objects' constraints are relationships between entities of an object model - entities here are objects, classes, attributes, links and associations.

Constraints exist in both inside and outside objects. An important characteristic of knowledge object modeling is that users can be given explicit control of the object hierarchy to customize the system to their particular needs, which includes letting users select among different methods for a particular task such as displaying the results.

Chapter 3 The Semantic Network

In the construction of a dictionary, we were presented with several difficulties due to the complex nature of biological data. These problems include multiple names for the same entity, the dependency of the biological state in which the function is taking place, and multiple functions for the same protein. To overcome these difficulties and to add additional functionality to our digital library, we have developed our dictionary as a semantic network.

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) which 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.

3.1 The UMLS Semantic Network

Our semantic network is 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. We will also have a category for medical findings including symptoms and laboratory results.

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.

3.2 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).

3.3 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 (*figure 4 and figure 5*).

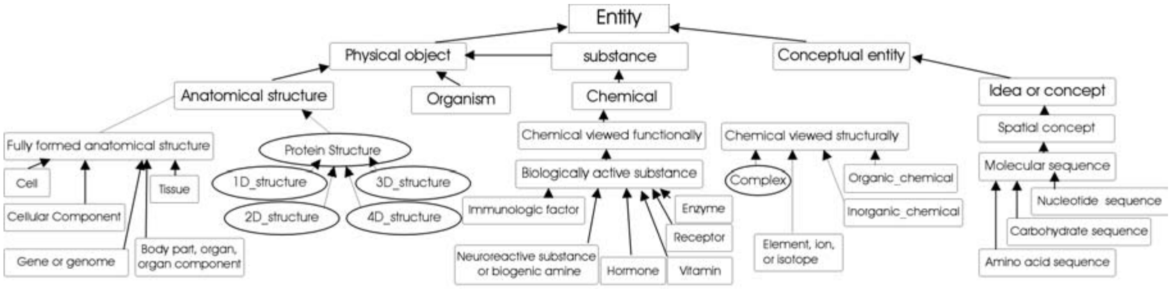


Figure 4. A simplified hierarchy showing a portion of our "entity" type.

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.

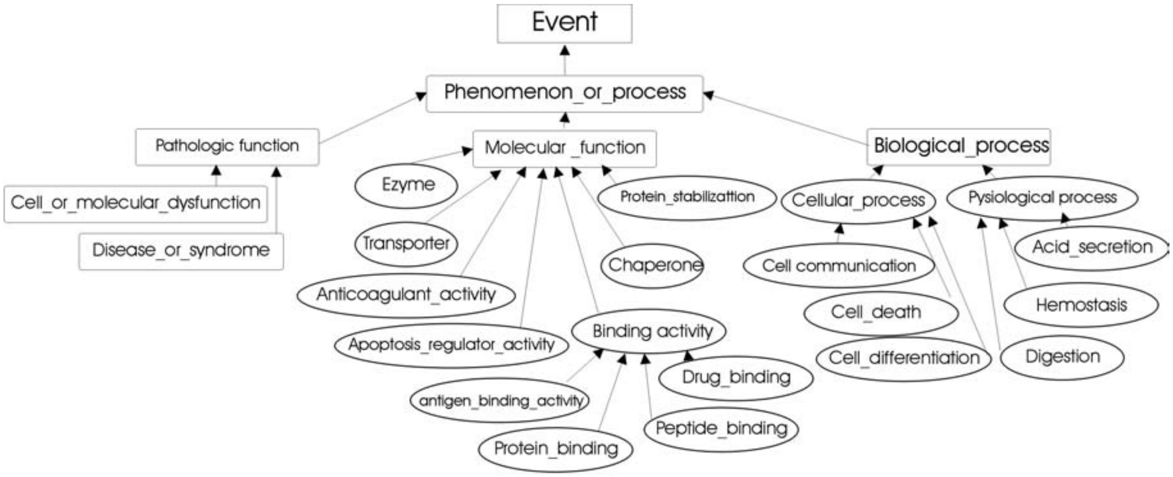


Figure 5. Semantic types can also represent events.

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.

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.

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 (*figure 6*).

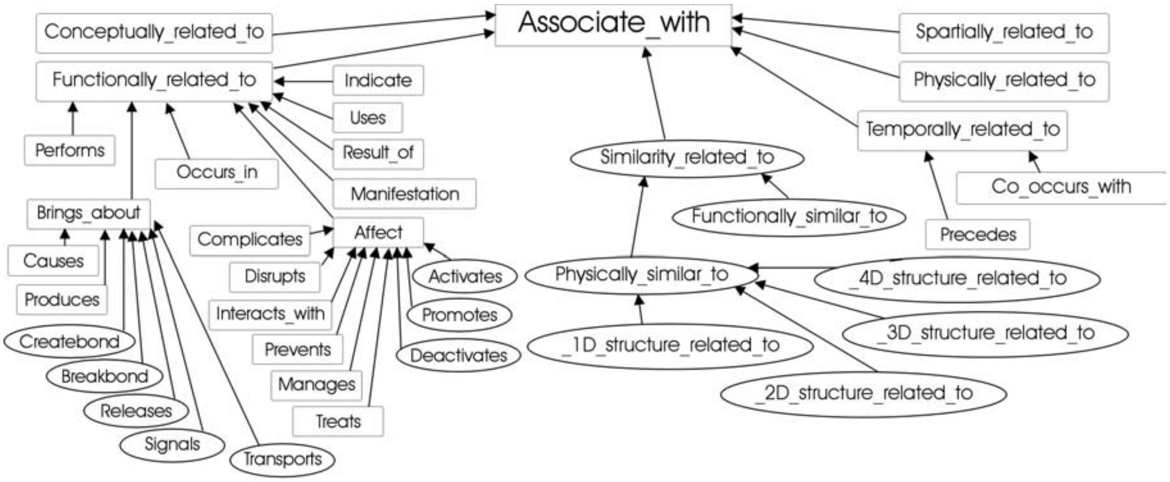


Figure 6. Semantic relationships tie the semantic net together.

In addition to the "is-a" relationships that represent a hierarchical structure, we also have "associate_with" relationships to represent the many non-hierarchical relationships 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.

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 the OMIM (Online Mendelian Inheritance in Man) database [OMIM]. This database of genetic disorders in man if 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 since these pages are dynamic and the underlying database is not accessible to the public. 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.

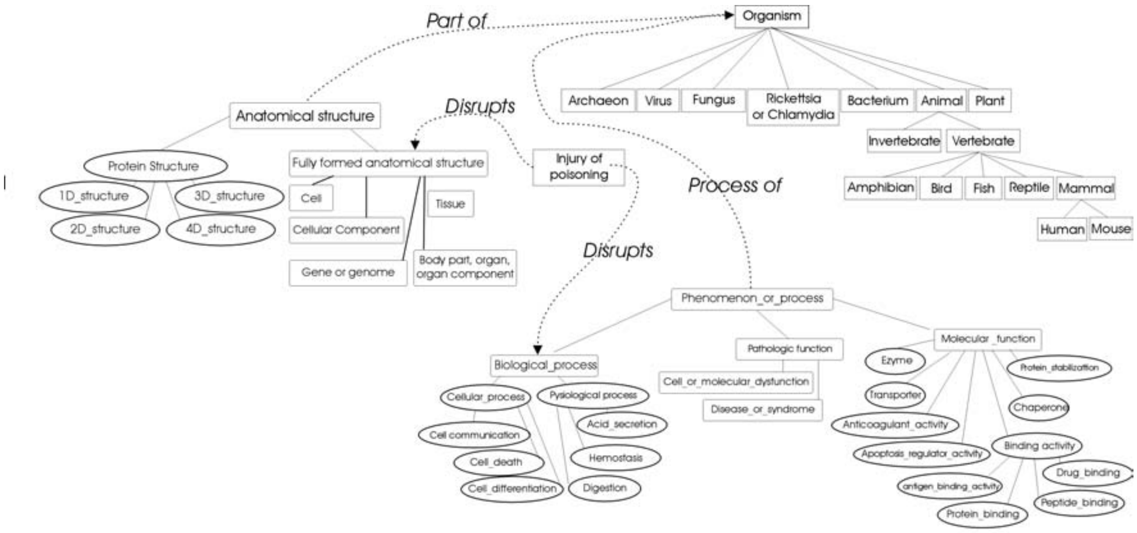


Figure 7. Shown here is a partial schema of the overall semantic network.

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.

Chapter 4 User Profiling and Recommendation Agents

4.1 Collecting User Information

Capturing user preferences can be a difficult task. Simply asking the users what they want can be obtrusive and error prone. In fact, the user might not even know what they really want. On the other hand, monitoring user behavior may be unobtrusive but can also be computationally time consuming and discovering meaningful patterns is difficult. Yet capturing user information is critical for the recommender system of our digital library.

User profiling methods can be classified as either knowledge based or behavior based. Knowledge based methods employ questionnaires or interviews to dynamically match users to one of a number of different static models of users. Behavior based approaches seek to capture the user's behavior and apply machine learning techniques to discover useful patterns in the behavior. This approach will need to log the user's behavior in some manner. Kobsa [Kobsa 93] provides a good survey of user modeling techniques.

The user profiling employed by recommendation agents is primarily behavior based. With most recommendation agents a binary, two-class model that represents what a user likes or dislikes is used. Machine learning techniques are then used to discover meaningful information about the user. In our system this meaningful information is in the form of rules. The recommendation agent will then use these rules to recommend items that she may be interested in.

The user knowledge can be collected either implicitly or explicitly. Implicit knowledge acquisition would be the preferred mechanism since it has little or no impact on the user's normal work. Analyzing the click stream as a user navigates through our system might provide one method for collect this information in an unobtrusive manner. This type of knowledge acquisition requires some degree of interpretation to understand the user's real interests. This is an inherently error prone process. How do we, for instance, determine if a user is lingering on one item because they are truly interested or if they were interrupted while navigating the site?

Explicit knowledge acquisition, on the other hand, requires the user to interrupt their normal work to provide feedback. This may be undesirable, but will generally provide the system with high confidence information since the user themselves provides the information. This feedback is most often in the form of a questionnaire on the relevance, interest and quality of an item. It may also come in the form of programming where the user is asked to create filter rules either visually or via a programming language.

Our system utilizes a combination of these different systems. When the user performs a search on our system, they may mark items as interesting. These items are then saved in the user profile. At any time the user may choose to use this profile to generate new rules. Due to the imprecision of this method, there is a third step where the user provides additional feedback on which rules to add to the profile.

4.2 The Weka Data Mining Package

The Weka machine learning libraries [Witten 2000] are an open source collection of data mining programs implemented in Java and issued under the GNU General Public License. Since they are open source, the algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. This system is available free of charge and provides tools for a good number of the machine learning algorithms that have been developed to date.

Data that is used in this system must conform to a specific text format called a ARFF file. This format requires each attribute to be declared with an "@attribute" tag. The format also expects a name for each attribute and a set of allowable values. This could be real numbers, but it is most often a set of classifications. The data section of the file is denoted with an "@data" tag and the data is entered in a comma-delimitated fashion matching the order of the @attribute tags.

Our system creates this ARFF file dynamically based on the user profile and knowledge information contained in each object in the library. The attributes are all of the keywords used to describe the items in our digital library. Each of these attributes is a Boolean, true or false attribute. The final attribute is the class for the item. For the recommendation agent, this class is either a yes or a no indicating the interest the user has for this item. For data section of the ARFF file, each item in our library is listed with the set of attributes pertaining to which keywords are associated with the item and a yes or no depending on the users prior interest to this item.

4.3 The J48 Decision Tree Algorithm

Our system builds user rules with the C4.5 tree algorithm developed by Ross Quinlan [Quinlan]. The Weka implementation of the C4.5 algorithm is contained in the J48 package. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm, summarized in Figure 8, is a version of ID3, a well known decision tree induction algorithm.

4.3.1 The Basic Strategy for Decision Trees

The basic strategy for building decisions trees can be described as follows:

- The tree starts as a single node representing the training samples (step 1).

- If the samples are all of the same class, then the node becomes a leaf and is labeled with that class (steps 2 and 3).
- Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes (step 6). This attribute becomes the "test" or "decision" attribute at the node (step 7). In this version of the algorithm, all attributes are categorical, that is, discrete-valued. Continuous-valued attributes must be discretized.
- A branch is created for each known value of the test attribute, and the samples are partitioned accordingly (step 8-10).
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node's descendents (step 13).
- The recursive partitioning stops only when any one of the following conditions is true:

  (a) All samples for a given node belong to the same class (step 2 and 3), or

  (b) There are no remaining attributes on which the samples may be further partitioned (step 4). In this case, majority voting is employed (step 5). This involves converting the given node into a leaf and labeling it with the class in majority among the samples. Alternatively, the class distribution of the node samples may be stored.

  (c) There are no samples for the branch *test-attribute* = a_i (step 11). In this case, a leaf is created with the majority class in samples (step 12)

4.3.2 The Attribute Selection Measure

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such and information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for i = 1,…,m). Let s_i be the number of samples of S in class C_i. The expected information needed to classify a given sample is given by

$$I(s_1, s_2, \ldots, s_m) = \sum_{i=1}^{m} p_i \log_2(p_i),$$

Where p_i is the probability that an arbitrary sample belongs to class c_i and is estimated by s_i/s. Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have v distinct balues, { a_1, a_2,…, a_v}. Attribute A can be used to partition S into v subsets, {s_1, s_2,…, s_v}, where s_j contains those samples in S that have value a_j of A. If A were selected as the test attribute (i.e., the best attribute for splitting), then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} (\frac{s_{1j} + ... + s_{mj}}{s}) I(s_{1j} + ... + s_{m}).$$

The term $(\frac{s_{1j} + ... + s_{mj}}{s})$ acts as the weight of the jth subset and is the number of samples in the subset (i.e., having a_j of A) divided by the total number of samples in S. The smaller the entropy value, the greater the purity of the subset partitions. Note that for a given subset s_j,

$$I(s_{1j}, s_{2j}, ..., s_{mj}) = -\sum_{i=1}^{\mathrm{m}} p_{ij} \log_2 (p_{ij}),$$

where $p_{ij} = \frac{s_{ij}}{|s_j|}$ and is the probability that a sample in S_j belongs to class C_i. The encoding information that would be gained by branching on A is

$$\text{Gain}(A) = I(s_{1j}, s_{2j}, \ldots, s_{mj}) - E(A).$$

In other words, Gain(A) is the expected reduction in entropy caused by knowing the value of attribute A.

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly. The ID3 algorithm is the basis for most decision tree algorithm (*figure 8*).

```
Algorithm: Generate_decision_tree. Generate a decision tree from the given training data.
Input: The training samples, samples, represented by discrete-valued attributes; the set of candidate attributes, attribute-list.
Output: A decision tree.
Method:

create a node N;
if samples are all of the same class, C, then
    return N as a leaf node labeled with the class C;
if attribute-list is empty then
    return N as a leaf-node labeled with the most common class in samples; //majority voting
select test-attribute, the attribute among attribute-list with the highest information gain;
label node N with test-attribute;
for each known value $a_i$ of test-attribute //partition the samples
    grow a branch from node N for the condition test-attribute = $a_i$;
    let $s_i$ be the set of samples in samples for which test-attribute = $a_i$; //a partion
    if $s_i$ is empty then
        attach a leaf labeled with the most common class in samples;
    else attach the node returned by Generate_decision_tree($s_i$, attribute-list - test-attribute);
```

Figure 8. The ID3 decision tree algorithm.

4.3.3 Extracting Classification Rules from Decision Trees

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule atecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large.

Chapter 5 System Architecture

There are many factors involved in determining the architecture of a digital library. In making this decision, one must determine how the system will be used and who the users will be. We want the system to be robust and scalable, but we also have to face the reality of a limited budget. We also want the system to be available to users throughout the UVM campus and also users from other institutions. The system must also combine several different agents together in a seamless manner, some of which may be difficult to modify.

Based on an analysis of the use of a system such as ours, we decided that a web application would be the ideal architecture for our digital library. Since the Weka data mining package was written in Java, we decided on a J2EE application server. We are using the Tomcat server to host our web application. This server is a free, open-source system available under the GNU Public License and is the reference implementation of the servlet 2.3 and JSP 1.2 specifications. Tomcat is quite powerful as either a stand-alone web server or embedded within an Apache server.

Tomcat can recognize standard HTML files, Java Server Pages (JSP) or Java servlets. Servlets are Java technology's answer to Common Gateway Interface (CGI) programming. They are programs that run on a Web server, acting as a middle layer between a request coming from a Web browser or other HTTP client and databases or applications on the HTTP server. It can be argued that Java servlets are more efficient, easier to use, more powerful, more portable, safer and cheaper than traditional CGI and other technologies. Java Server Pages allow one to include Java code inside of an HTML page. This provides the author close control over the design of the page.

5.1 Model View Controller (MVC) Architecture Design.

Several problems can arise when applications contain a mixture of data access code, business logic code, and presentation code. Such applications are difficult to maintain, because interdependencies between all of the components cause strong ripple effects whenever a change is made anywhere. High coupling makes classes difficult or impossible to reuse because they depend on so many other classes. Adding new data views often requires re-implementing or cutting and pasting business logic code, which then requires maintenance in multiple places. Data access code suffers from the same problem, being cut and pasted among business logic methods.

The Model-View-Controller design pattern solves these problems by decoupling data access, business logic, and data presentation and user interaction. As the name suggests, there are three main components of the MVC design. The model, view and controller are described as follows:

- **Model** - The model represents enterprise data and the business rules that govern access to and updates of this data. Often the model serves as a software approximation to a real-world process, so simple real-world modeling techniques apply when defining the model.

- **View** -The view renders the contents of a model. It accesses enterprise data through the model and specifies how that data should be presented. It is the view's responsibility to maintain consistency in its presentation when the model changes. This can be achieved by using a push model, where the view registers itself with the model for change notifications, or a pull model, where the view is responsible for calling the model when it needs to retrieve the most current data.

- **Controller** - The controller translates interactions with the view into actions to be performed by the model. In a stand-alone GUI client, user interactions could be button clicks or menu selections, whereas in a Web application, they appear as GET and POST HTTP requests. The actions performed by the model include activating business processes or changing the state of the model. Based on the user interactions and the outcome of the model actions, the controller responds by selecting an appropriate view.

The separation of model and view allows multiple views to use the same enterprise model. Consequently, an enterprise application's model components are easier to implement, test, and maintain, since all access to the model goes through these components. In addition, the MVC architecture promotes reuse of model components. To support a new type of client, you simply write a view and some controller logic and wire them into the existing enterprise application. This type of design pattern will introduce some extra classes due to the separation of model, view, and controller. However, the benefits of code of re-use and ease of maintenance more than make up for this increase in complexity.

5.2 Components of Our Digital Library

The primary feature that sets our system apart from other systems is the recommendation agent. This agent will generate rules about the users and learn about the users' interests and preferences. The rules will be refined and improved through two learning processes: interactive incremental learning and silent incremental learning. Our system will first learn about a user's areas of interest by analyzing the user's declared interest topics and the user's visit records, and then assist the user in retrieving the right information. The user profile is composed of a set of biological terminologies coming from the knowledge structure and the dictionary. Interactive incremental learning will function in cycles that interact with the user. The system will prompt the user with a set of related documents which are likely of the user's interest, and ask for feedback on the level of interest in each of these documents. Considering the feedback, the system will make changes to its search and selection heuristics and improve its performance.

Our system will work differently from search engines and other kinds of agents like WebWatcher [Joachims et al 1995] and [World Wide Web Worm] that help the user on the global Web. First, through incremental learning of the user's characteristics or interest areas, the system will become an assistant to the user in retrieving relevant information. Second, our library will have the potential to reduce user accessing and retrieval time, by displaying a list of changes that have been made since the user's last visit. Finally, system can be easily adopted for other digital libraries. This can be accomplished by adding a different knowledge source for the dictionary.

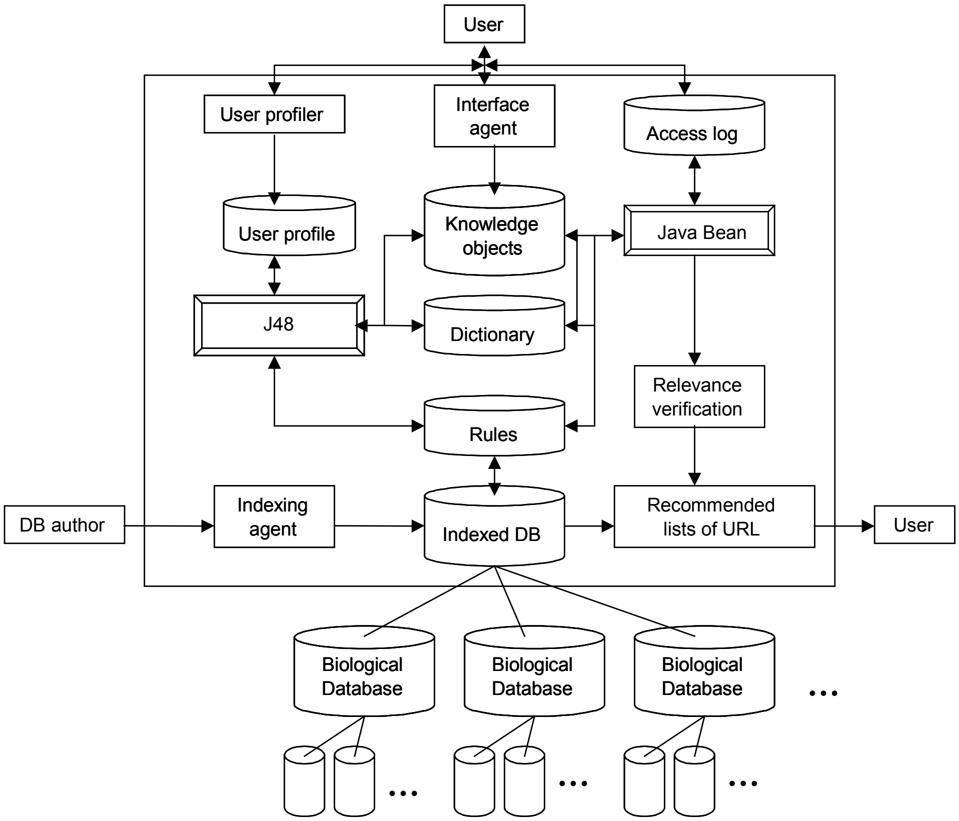


Figure 9. System architecture design.

An overall schema of our digital library system. The J48 induction engine will be the "brain" of the digital library. 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.

Figure 9 shows the design of our digital library system structure which is a modification of an existing prototype [Ngu & Wu 1997]. The relevance verification agent (RA) and document updates agent (DA) have not yet been implemented. This will be the work of future projects. We will therefore concentrate on the components of the system that have been implemented and relate to this project.

5.2.1 Access Log

Most web sites allow global user access and have logging facilities in place [Pitkow & Bharat 1994] to record users' access details. The access log of a web site records all web transaction/request services by the web server. The three main elements for each record are: the machine name with its Internet address from which the access is performed, the date/time of access, and the Web page being accessed. Future versions of our digital library will make use of these logs in the creation of an update agent.

5.2.2 Dictionary

The dictionary has been a key focus of our research. The dictionary is implemented as a semantic network of biological terminology and represents the knowledge source for our library. It will be used by various agents including the recommendation agent. All keywords that describe items in our digital library must come from this dictionary, however the use of concept classes allows for broad leeway in these keywords. Chapter 3 of this thesis discusses the design and implementation of the dictionary in detail.

As described in Chapter 3, the semantic network not only contains the terminology used in the keywords, but also their relationships. This will allow for the implementation of several intelligent agents to provide additional functionality in future versions of the library. In this version of the library, the primary relationship is at the concept class level. For each keyword, we search the metathesaurus for the preferred term and substitute this term for the keyword. This will allow the system to find not only those items that have the same keywords, but also those that share the same meaning. This is a large improvement over the standard keyword searching employed by most library systems. We have chosen to do this keyword substitution at runtime and not by altering the keywords associated with an item. Although this will make the system a little slower, it will allow for changes in the semantic network to propagate throughout the application and will preserve the original information provided with each item in the library.

In our system we have implemented the semantic network in a relational database. This database includes tables for both the semantic network and for the metathesaurus of terms at the concept class level. Each term has a unique identifier and a concept identifier. All terms that have the same concept identifier have similar meanings and make up the concept class for those terms. Each concept class has one and only one term which is the preferred term for that concept class and is stored in a table of preferred terms indexed by the unique concept identifiers. A separate table links the concept identifiers to a list of different relationship identifiers. These relationship identifiers are stored in a table structure similar to that of the concepts.

5.2.3 Weka Data Mining Package

The Weka data mining library will serve as the “brain” of the discovery agent. It will take two input sets of documents; one set the user has seen and selected as interesting, and the other the user has not visited or has not found to be of interests. 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.

5.2.4 Indexed Database

The indexed database will have an entry/index for each biological item in the digital library. We are using the MySQL database for the semantic network described, the user profile, and items in our library. Each entry in the database will have a pointer to the corresponding item, with a set of keywords from the dictionary to index the biological item, a date and time to show when the item was last modified, and a format indicator. The entries can be nested to allow for compound documents to be searched. In the future, we plan to employ the Resource Description Framework (RDF) [W3C] to describe keyword properties of documents and compound documents, and integrate a commercial object-oriented database system for the indexed database, in order to manipulate large collections of documents effectively [Schatz & Chen 1996].

5.2.5 Indexing Agent

Future versions of the library will employ this agent to traverse the digital library and index all relevant biological information according to the dictionary and store the results in the indexed database. It will incrementally refresh the indexes and automatically update indexing when document updates are forwarded by the database authors or administrators. How to generate and maintain the list of keywords for each document will be a crucial issue in future versions of this project.

5.2.6 Interface Agent

The interface, i.e. what the user sees, will be developed using the “view” component of the MVC design and is implemented with Java Server Pages. The goal is to allow the user to interact easily with the system. The agent will provide the user with functions for navigation of the library, evaluation of retrieved documents, setting up user preferences, and a help system with hyperlinks using the semantic links between the keywords in the dictionary. We also will have a facility on the digital library interface for biological authors to submit relevant materials directly in electronic format.

5.2.7 User Profiles

Gathering the information which is thought to be of interest to the user is one of the most difficult tasks in digital library system development. A user profile will consist of the user's account details, areas of interest, access history, and the rules generated by Weka. Information about a user’s interests describes user’s information needs in terms of information types and types of contents, and short-term and long-term interests. Since the domain is restricted to a very specific area of research, it would be useful in constructing user profiles to have information about each user. This personal data is a collection of user’s personal identification data, such as user’s name, status (student, researcher, etc.), user’s current field of research, etc. More sophisticated information about user’s interests are established by referring to both the knowledge structure and the dictionary. Each user can maintain more than one user profile, which enables the user to work on multiple subjects. One of them is assigned as a default profile that is running without specific selection of choice. To make this possible, we will provide an interface in which the user can switch the preference. Whenever the user switches the profile, the session is closed and new session begins with the new profile.

5.2.8 User Machine with Web Browser.

A Web user can access our digital library through a Java enabled Web browser. Most modern Web Browsers such as Microsoft Internet Explorer, Mozilla, and Netscape support Java.

5.3 Indexing Biological Information and Reporting Updates.

How to generate and maintain the list of keywords for individual biological information is a crucial issue to the project, because the keywords are the indicator of information content to which the digital library is to have access, and if relevant keywords are missing, the system will not be able to assist the user. We assume that all the information in the databases is completely indexed by the database authors and database administrators. Any updates made by the database authors and database administrators are notified and sent to the indexed databases to appropriately update the indexing. At the same time, the updates are consulted with the user profiles and corresponding users for update notification are determined. During this decision making process, keywords in the user profiles are compared with those in the keyword dictionary and semantic links between these keywords. The relevance verification agent (RA) in Figure 9 will make use of the indexing facility. If a new document cannot be indexed properly by the keywords in the dictionary, the document will be considered irrelevant to the digital library.

Chapter 6 Comparison with Related Work

A crucial factor in sharing of biological information among users in a consistent manner is whether a standardized framework exists and unambiguous information can be expressed and communicated. One existing method to capture the concepts and relationships between them is ontology. Ontologies have been used in Artificial Intelligence to describe a variety of domains. Initial endeavors to include ontology for biology can be found in [Schulze-Kremer 1998] and [ISMB 1998].

A dictionary for biology is one possibility for biological knowledge representation that reflects a specific view of the data. This ontology has been adopted in the TAMBIS project and the Gene Ontology Consortium's controlled vocabulary system. [Baker et al 1998] (*figure 10*). Among three models, the biological concept model, a knowledge base of biological terminology, is central to its architecture. This is basically used to drive query formulation and facilitate source integration in their system. However, in our proposal we will have an additional use for this, which is, building and maintaining user profiles. Since the dictionary provides a scheme that can reason about the relationships between terms and their components, the terminologies in the dictionary can indicate the possible interests of the user in multiple levels of the subject category.

Assisting Web users by identifying their areas of interest has attracted the attention of quite a few recent research efforts. Several recent research projects ([Balabonovic & Shoham 1995] and [Yan et al 1996], WebMate [Chen & Sycara 1998], Three-Descriptor

Represenation [Widyantoro et al 2000] and Web Learner [Pazzani et al 1995]) also share similar ideas.

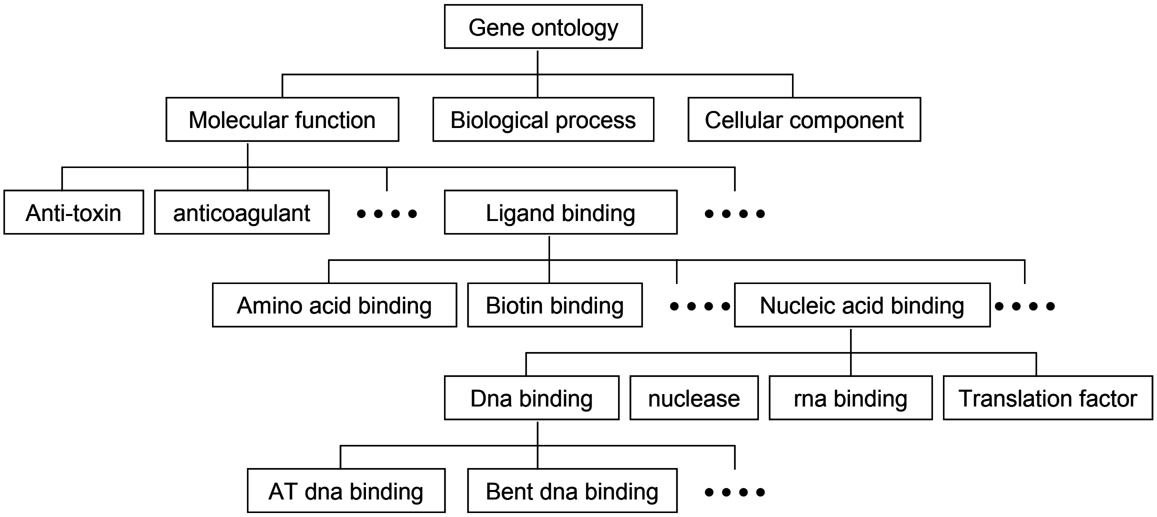


Figure 10. Gene Ontology Hirerarchy

[Balabonovic & Shoham 1995] developed a system that helps a Web user discover new sites that are of the user's interest. The system presents the user every day with a selection of Web pages that it thinks the user would find interesting. The user evaluates these Web pages and provides feedback for the system. The user's areas of interest are represented in the form of (keyword, weight) pairs, and each Web page is represented as a vector of weights for the keywords in a vector space[1]. From the user's feedback, the system knows more about the user's areas of interest and better serves the users in the future. If the

[1] The vector space approach is one of the most promising paradigms and the best-known technique in information retrieval.

user's feedback on a particular Web page is positive, the weights for relevant keywords of the Web page are increased, otherwise they are decreased. This system adds learning facilities to existing search engines, and as a global Web search agent does not avoid the general problems associated with search engines and Web robots. In addition, compared to our system, the (keyword, weight) pairs used in this system cannot represent logical relations between different keywords, such as (``data mining'' AND ``Internet'') OR ``rule induction''. This type of logical expressions will be the starting point for knowledge representation and data mining in our system.

[Yan et al 1996] investigated a way to record and learn user access patterns in the area of designing on-line catalogues for electronic commerce. This approach identifies and categorizes user access patterns using unsupervised clustering techniques. *User access logs* are used to discover clusters of users that access similar pages. When a user comes, the system first identifies the user's pattern, and then dynamically reorganizes itself to suit the user by putting similar pages together. An (item, weight) vector, similar to the (keyword, weight) vector used to represent each Web page in [Balabonovic & Shoham 1995], is used in [Yan et al 1996] to represent a user's access pattern. The system views each Web page as an item, and the weight of a user on the item is the number of times the user has accessed the Web page. This system does not use semantic information (such as areas of interest) to model user interests, but just actual visits. Also, it does not aim to provide users with newly created or updated Web pages when they visit the same Web site again. This is a significant difference in design between this system and our ours.

WebWatcher [Armstrong et al 1995] is an agent that helps the user in an interactive mode by suggesting pages relevant to the current page the user is browsing. It learns by observing the user's feedback to the suggested pages, and it can guide the user to find a particular target page. A user can specify their areas by providing a set of keywords when they enter WebWatcher, mark a page as interesting after reading it, and leave the system at any time by telling whether the search process was successful or not. WebWatcher creates and keeps a log file for each user and from the user's areas of interest and the “interesting” pages they have visited, it highlights hyperlinks on the current page and adds new hyperlinks to the current page. WebWatcher is basically a search engine, and therefore does not avoid the general problems associated with search engines and Web robots. Although it has been extended to act as a tour guide [Joachims et al 1997], it does not support incremental exploration of all relevant, newly created and updated pages at a local site.

Three-Descriptor Represenation [Widyantoro et al 2000] learns the areas that are of interest to a user, by recording the user's browsing behaviour. It performs some tasks at idle times (when the user is reading a document and is not browsing). These tasks include looking for more documents that are related to the user's interest or might be relevant to future requests. Different from WebWatcher, Three-Descriptor Representation is a user interface that has no predefined search goals, but it assumes persistence of interest, i.e., when the user indicates interests by following a hyperlink or performing a search with a keyword, their interests in the keyword topic rarely end with the returning of the search results. There are no specific learning facilities in Three-Descriptor Represenation (but just a set of heuristics like the persistence of interest plus a best-first search), and therefore it does not perform incremental learning as our system will.

Web Learner [Pazzani et al 1995] is similar to our system in that it learns about what a user is interested in and decides what new Web pages might interest the user. However, Web Learner generates keywords (called a feature vector) automatically from pages on the global Web, and does not provide facilities for incremental learning. Furthermore, none of the extensions mentioned in Issue II(c) of Section D.2.2 have been addressed in Web Learner.

Our localized digital library agent will start with the same idea of assisting Web users by learning and identifying their areas of interest. However, agent will work with a centralized digital library server which contains indexes to Web pages on the Web by using a keyword dictionary local to the digital library. Further, based on the indexing of the Web pages on and linked to the digital library server, our system will support interactive and incremental learning. The rules with logical conditions in our system will be more powerful than the (keyword, weight) pairs used in some existing systems in representing users' areas of interest.

Our system will be different from existing search engines and robots on the World Wide Web. It does not traverse the global Web, but acts as a housekeeper for a centralized digital library server and as a helper for the user who visits the digital library to find relevant information, with particular attention to the newly developed and modified documents in the digital library.

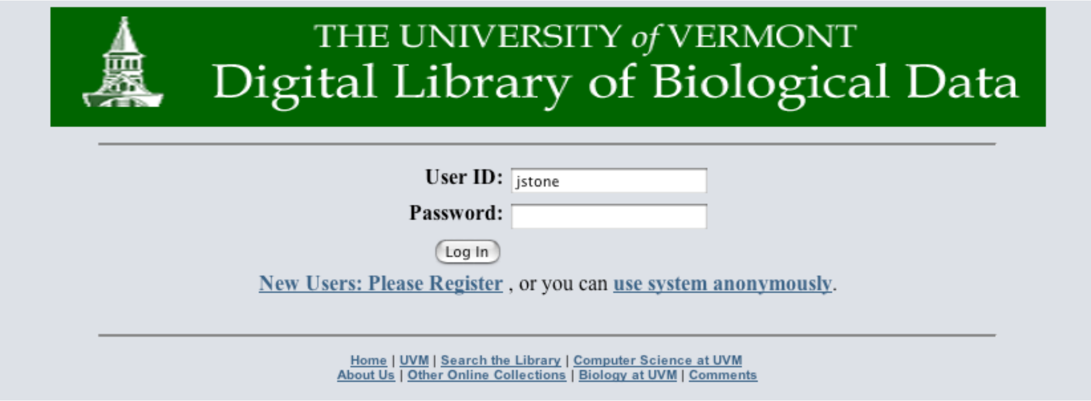
Chapter 7 Results and Future Work

7.1 The Content of Our Library

To illustrate the digital library system developed in this project, we have built a specialized digital library for gene analysis based on the work of my co-advisor, Dr. Greenblatt. Databases of literature, sequence, structure, and others will be used, including GenBank, dbSNP, the Molecular Modeling Database (MMDB), PubMed, Online Mendelian Inheritance in Man (OMIM), the p53 mutation database (http://www.iarc.fr/p53/), and a p16 mutation database compiled by Dr. Greenblatt [Murphy et al]. We will demonstrate the flexibility of our system by including multiple types of information including text, sequence information, 3-D protein structure information and even entire databases. All items were entered into the system using the add items interface in the administrative window.

7.2 The Digital Library Interface

Our system (http://www.cs.uvm.edu:9180/library) has two entry points: one for general users and one for administrators. Figure 11 below shows the entry to the site for general users. From entry point, new users create an account by clicking on the "New Users: Please Register" or they can enter the system anonymously. Anonymous users will not be able to use the recommendation agents until they open an account on the system. They will however have the ability to use the search capabilities of the site and can browse the site as they wish.



THE UNIVERSITY *of* VERMONT
Digital Library of Biological Data

User ID: jstone
Password:

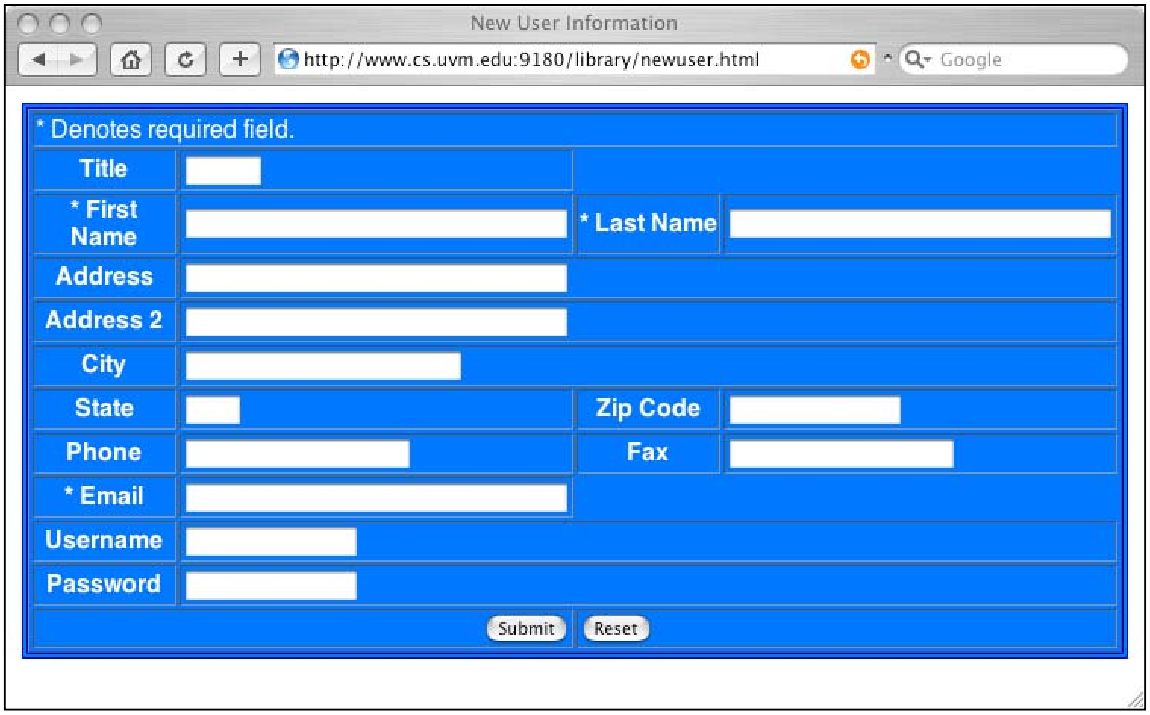
New Users: Please Register , or you can use system anonymously.

Log In
Home | UVM | Search the Library | Computer Science at UVM
About Us | Other Online Collections | Biology at UVM | Comments

Figure 11. The entry page for general users.

Users who register with the system will fill out a short questionnaire that collects demographic information as well as a username and password for the system (*figure 12*). Once the user is registered with the system, they will have full access to all non-administrative functions of the library. The system will create a user profile in XML format and will add the user to the database of library users.

After a user account has been created, the user can enter the digital library by providing their username and password. They then will find themselves on the main search page for the digital library. From this page they can access all other functions of the digital library including searching the library, browsing the semantic network, generating new rules based on their profiles and viewing items recommended by the system.



New User Information

http://www.cs.uvm.edu:9180/library/newuser.html

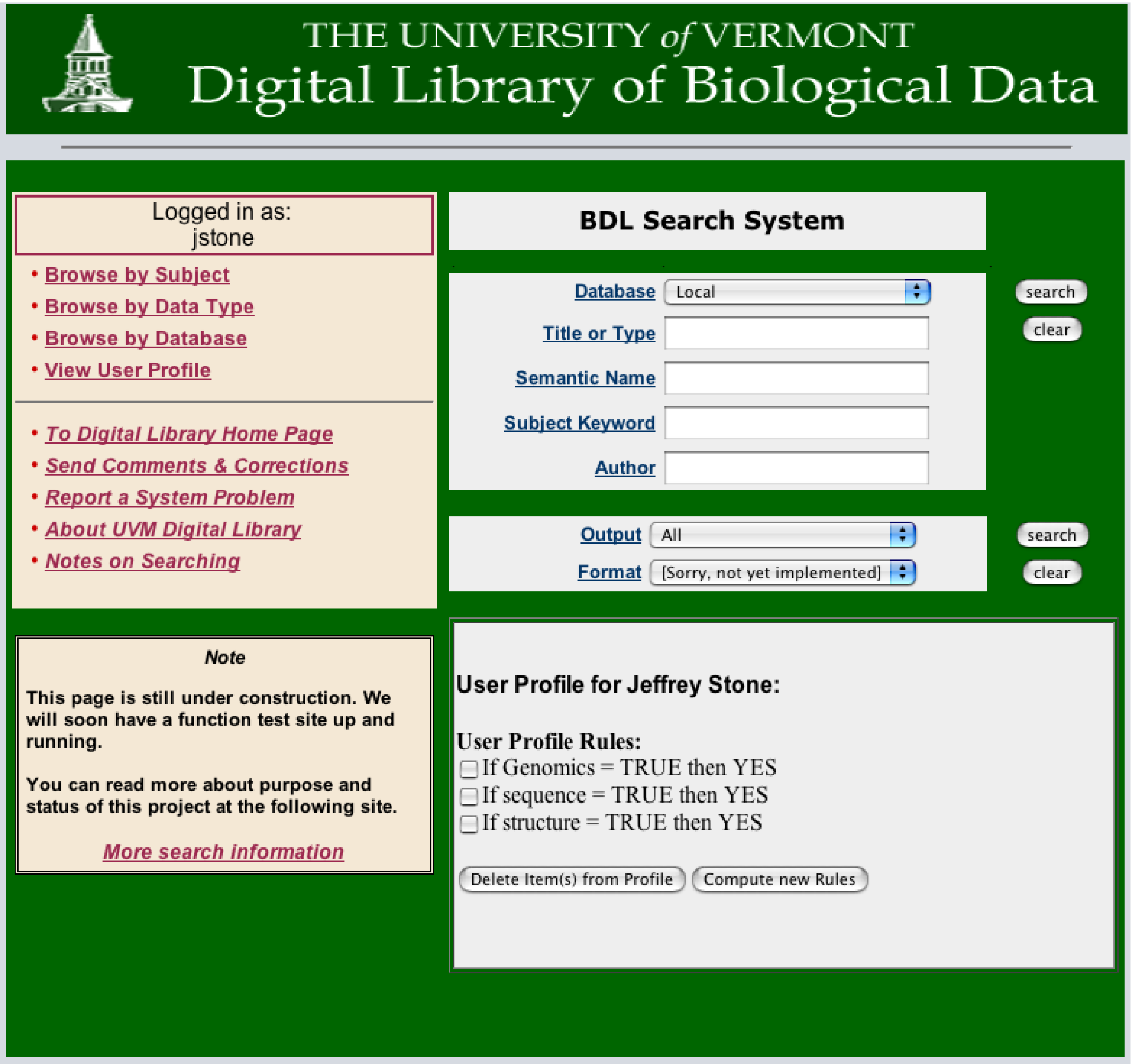
* Denotes required field.

Title			
* First Name		* Last Name	
Address			
Address 2			
City			
State		Zip Code	
Phone		Fax	
* Email			
Username			
Password			

Figure 12. New user questionnaire.

The left hand portion of the main page includes a number of links to information about the digital library and different functions for the library in general. These functions include browsing the library by data type, subject, or database. These functions print out to the screen all data in the site organized by the parameter selected. This functionality will become nearly useless in a large library system, but could be subdivided in future versions of the system to provide a clickable map of the data by utilizing the relationships in the semantic network.

This section also includes the function to view your user profile. This will open a window displaying your user profile in XML format. This is the file used by the intelligence agents of the system. It is also used to display your user profile at the bottom of the right hand side of the page (*figure 13*).



THE UNIVERSITY *of* VERMONT
Digital Library of Biological Data

Logged in as:
jstone

- Browse by Subject
- Browse by Data Type
- Browse by Database
- View User Profile

- *To Digital Library Home Page*
- *Send Comments & Corrections*
- *Report a System Problem*
- *About UVM Digital Library*
- *Notes on Searching*

BDL Search System

Database: Local
Title or Type:
Semantic Name:
Subject Keyword:
Author:

Output: All
Format: [Sorry, not yet implemented]

Note

This page is still under construction. We will soon have a function test site up and running.

You can read more about purpose and status of this project at the following site.

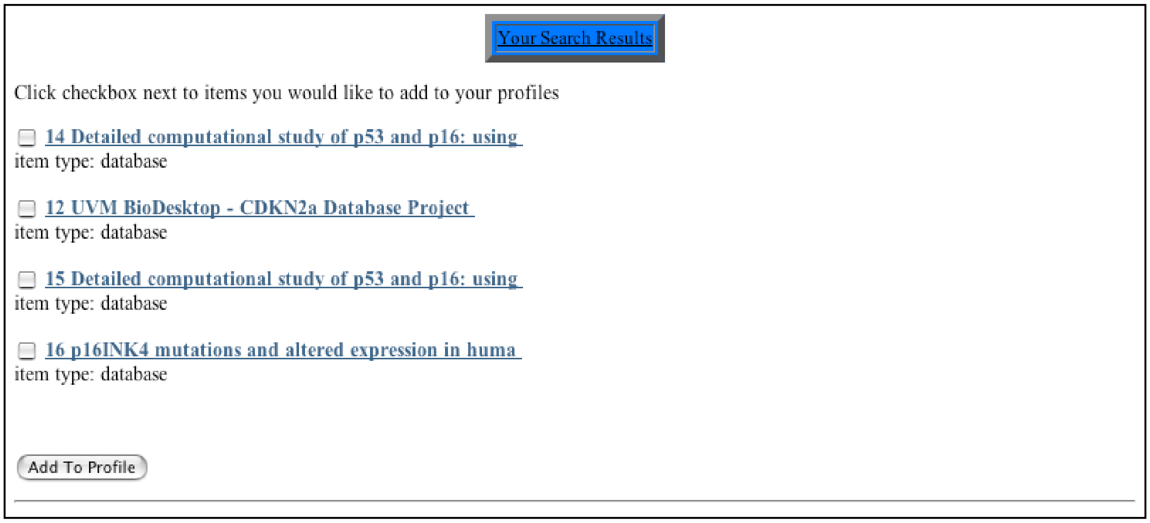
More search information

User Profile for Jeffrey Stone:

User Profile Rules:
□ If Genomics = TRUE then YES
□ If sequence = TRUE then YES
□ If structure = TRUE then YES

Figure 13. The main page for the library.

The upper right hand side of this page provides the search interface to the library. You can search the library by title, semantic type, subject keyword, or author. At the top of this section there is a selector for the database that you want to search. In addition to searching our library, you can use this interface to search several other databases including the very popular NCBI Entrez site. When searching off site databases, control will be transferred to the off site system in a new window. In addition to searching the off site database with the search criteria specified, your user profiles will also be used to recommend other items on those sites. This provides a valuable additional functionality over using those web sites directly.



Click checkbox next to items you would like to add to your profiles

□ 14 Detailed computational study of p53 and p16: using
item type: database

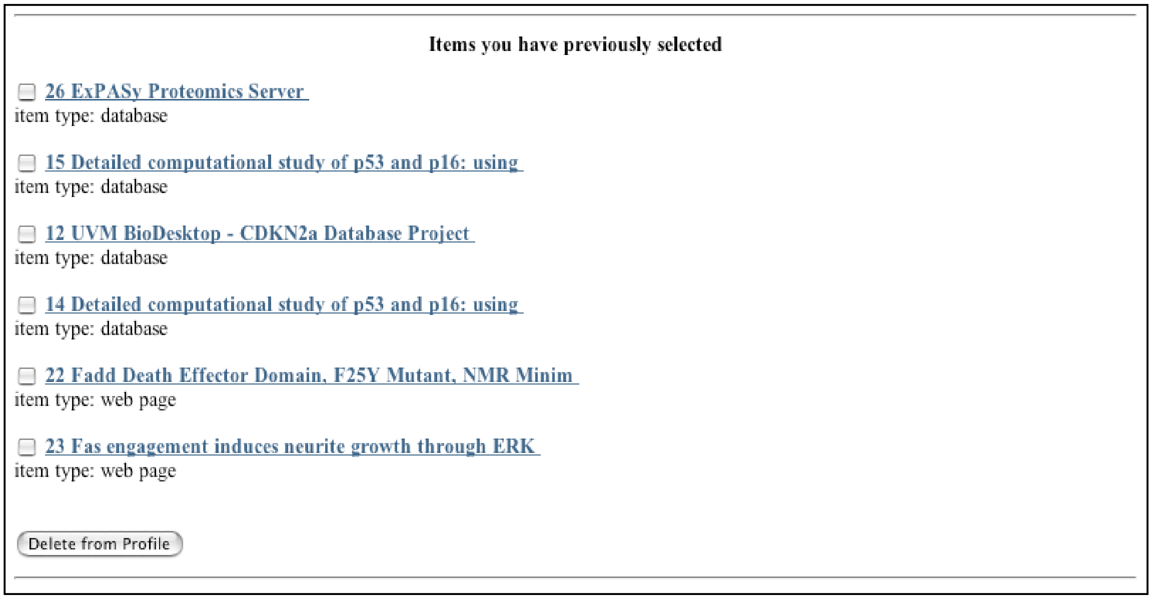
□ 12 UVM BioDesktop - CDKN2a Database Project
item type: database

□ 15 Detailed computational study of p53 and p16: using
item type: database

□ 16 p16INK4 mutations and altered expression in huma
item type: database

Figure 14. Search results for p16.

The results of a search are presented in a new separate window. The results will be shown as a list of web links to the items located on the Internet. Each item will also have a short description and a check box beside it to indicate if the item was helpful or not. Selecting an item will open up the web page for that item in a separate window. This allows you to open several items at a time and still have your result list available to refer back to. In addition to this format, you may also select an alternate output format such as XML or the ANS 1.1 format used by NCBI.



Items you have previously selected

□ 26 ExPASy Proteomics Server
item type: database

□ 15 Detailed computational study of p53 and p16: using
item type: database

□ 12 UVM BioDesktop - CDKN2a Database Project
item type: database

□ 14 Detailed computational study of p53 and p16: using
item type: database

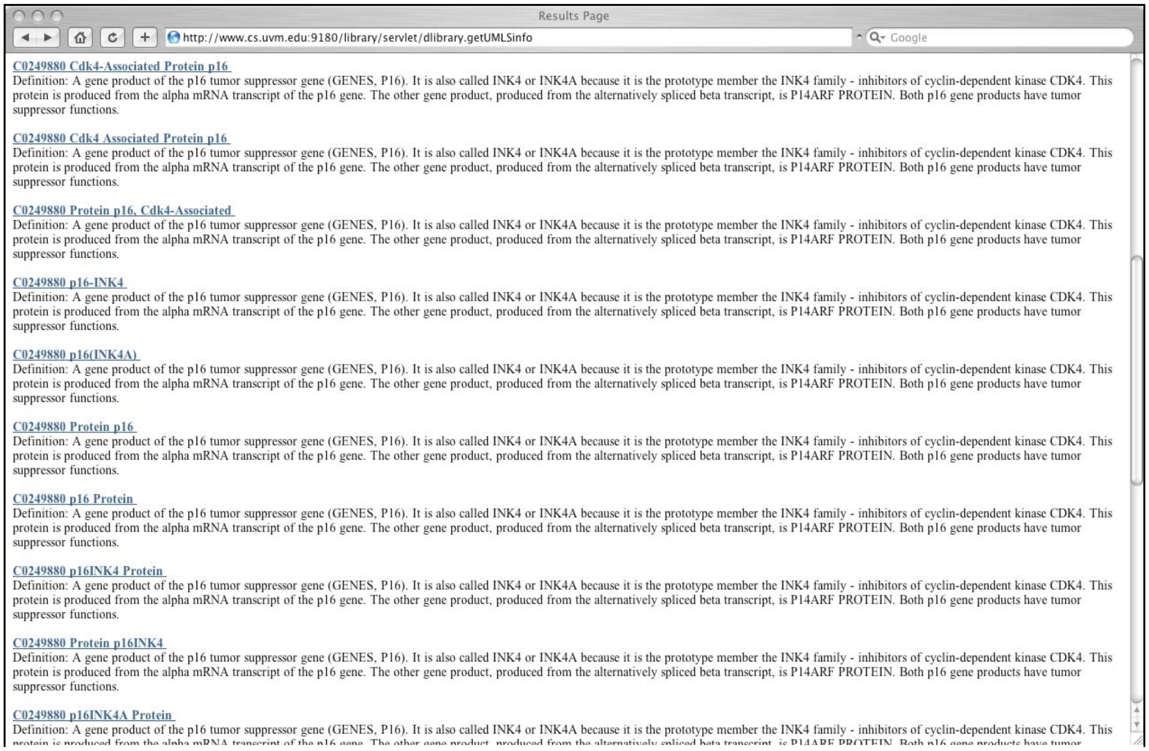
□ 22 Fadd Death Effector Domain, F25Y Mutant, NMR Minim
item type: web page

□ 23 Fas engagement induces neurite growth through ERK
item type: web page

Figure 15. Items previously selected are also displayed.

In addition to displaying the items found in your search, two other sections are displayed. The first of these is the listing of items previously selected as interesting. You may choose to delete any of these from you profile. Future versions of this system will also include an update agent that will notify the users of new information on a selected item. At the bottom of the page is the items that the system has recommended to you. This recommendation is based on the user rules in your profile.

In addition to searching by author, keyword or title, you may also search by semantic type. This is a distinct search type that is unlike the others. Figure 16 below shows an example of searching the semantic type for p16. This search will return all terms that are related to p16. This list can be quite exhaustive, but can helpful for finding items that are at the “tip of your tongue”.


Results Page
http://www.cs.uvm.edu:9180/library/servlet/dlibrary.getUMLSinfo
Google

C0249880 Cdk4-Associated Protein p16
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 Cdk4 Associated Protein p16
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 Protein p16, Cdk4-Associated
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 p16-INK4
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 p16(INK4A)
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 Protein p16
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 p16 Protein
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

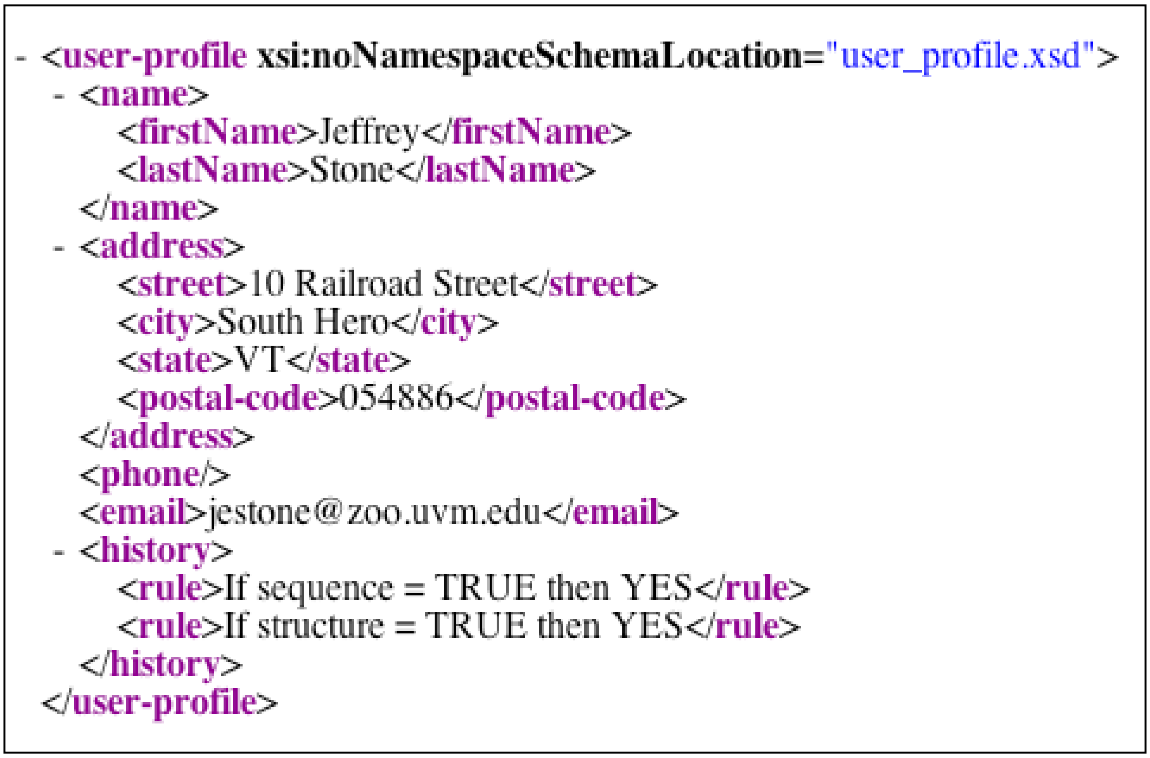
C0249880 p16INK4 Protein
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 Protein p16INK4
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor suppressor functions.

C0249880 p16INK4A Protein
Definition: A gene product of the p16 tumor suppressor gene (GENES, P16). It is also called INK4 or INK4A because it is the prototype member the INK4 family - inhibitors of cyclin-dependent kinase CDK4. This protein is produced from the alpha mRNA transcript of the p16 gene. The other gene product, produced from the alternatively spliced beta transcript, is P14ARF PROTEIN. Both p16 gene products have tumor

Figure 16. A partial display of the semantic search for "p16".

At the bottom of the main page is a window that provides the user interface for controlling their profile. This section uses a XSLT transformation to generate this section based on your XML formatted user profile. You can see your raw XML file by clicking on “View User Profile” on the upper left hand panel. XSLT transformations use this XML file to build a HTML formatted section with a list of your user rules and buttons for removing rules from your profile or computing new rules. When you compute new rules, the Weka J48 algorithm is used to generate a decision tree based on your profile and then uses this decision tree to generate the new rules.



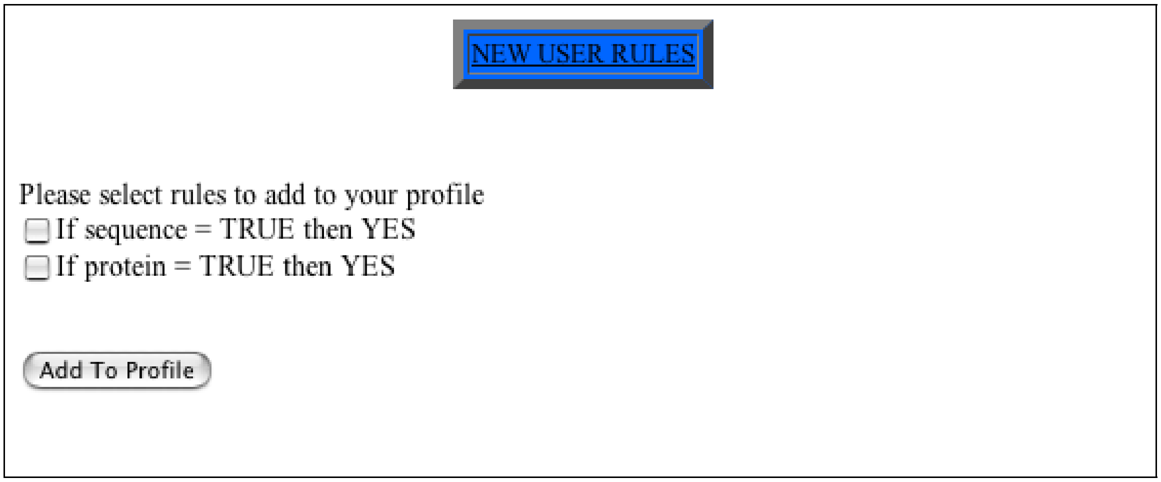
```
- <user-profile xsi:noNamespaceSchemaLocation="user_profile.xsd">
  - <name>
      <firstName>Jeffrey</firstName>
      <lastName>Stone</lastName>
    </name>
  - <address>
      <street>10 Railroad Street</street>
      <city>South Hero</city>
      <state>VT</state>
      <postal-code>054886</postal-code>
    </address>
    <phone/>
    <email>jestone@zoo.uvm.edu</email>
  - <history>
      <rule>If sequence = TRUE then YES</rule>
      <rule>If structure = TRUE then YES</rule>
    </history>
  </user-profile>
```

Figure 17. A sample user profile in raw XML format.

Since many researchers would be concerned by ad hoc generation of new rules, and because of the imprecision in the generation of these rules, user feedback is again requested. The list of rules found is displayed for the user to view. The user is then asked to select those rules that are of meaningful for their research (*figure 18*).

7.3 Future research directions.

This project is a work in progress. We are submitting a grant with these preliminary findings to support this project and to expand upon the system. The portion of the existing library that needs the most immediate attention is the database of items. This database is very small at this time. We hope that this system will grow in the very near future to provide a valuable asset not only to the UVM research community, but indeed the world. Some of the specific areas where future versions of this system may evolve are described below.


NEW USER RULES

Please select rules to add to your profile
□ If sequence = TRUE then YES
□ If protein = TRUE then YES

Figure 18. Generation of new user rules.

Users can select which rules they would like to add to their profile. In addition to simple rules as shown above, J48 can discover complex rules such as the conjunction of rules.

7.3.1 Sophisticated data entry for populating the library.

At this time, the administrative page for logging items into the library is html form page. This process is very tedious and error prone. We would like to create an agent for the automatic entry of these items. Potentially some of the emerging web services will provide some tools or method for entering this information directly from trusted sites.

7.3.2 Improvement in Semantic Searching.

The current method for searching the semantic network for relationships is quite slow. This is due in part to the large size of the semantic network database (roughly 3 million entries). Some effort will need to be spent on improving the performance of this semantic search. Perhaps there may be better ways of implementing the semantic network via java object creation in memory.

7.3.3 Emerging Semantic Web Technologies.

One of the more exciting areas of research is in the semantic web technologies. New means of annotating web resources are promising to revolutionize the way we use the Internet. The current internet is designed mostly in markup languages to format this information for human consumption. With a new focus on web services and XML technologies, researchers are looking into ways to create a web designed by machines for machines.

Semantic Web technologies such as DAML and OIL could potentially be used for the generation of our semantic net [Berners-Lee 2001]. This would allow this net to extend beyond our own system to include other semantic networks and web services by searching the descriptors such as Resource Description Frameworks (RDF) for these systems for common nodes in the networks.

7.3.4 Expanding the Ontology

The current ontology contains information primarily designed for describing molecular biologists. Expanding this ontology can be achieved by either adding new nodes to the semantic network, or by adding whole new semantic networks to the system. The modular design of our system allows one to plug in any semantic network that fits the general schema outlined in Chapter 3. This would allow our library to be used outside of its intended biological domain to any domain that one would want.

7.3.5 Additional User Agents

There are several additional user agents that could be developed for our system. Chief among these would be an update agent. This agent would notify the user of updates to items that fit their profile. The semantic network contains a valuable knowledge base that could be further exploited. An agent for navigating the semantic neighborhood of an item in a graphical map might also prove interesting.

Bibliography

- [Amato & Straccia 1999] G. Amato & U. Straccia, User Profile Modeling and Applications to Digital Libraries, *3rd European conference on digital libraries*, Paris, France, 1999.
- [Armstrong et al 1995] R. Armstrong, D. Freitag, T. Joachims and T. Mitchell, WebWatcher: A Learning Apprentice for the World Wide Web, *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, March 1995.
- [Baker et al 1999] P. Baker, C. Goble, S. Bechhofer, N. Paton, R. Stevens, A. Brass, An Ontology for Bioinformatics Applications, *Bioinformatics*, Vol. 15 No. 6. pp 510-520, 1999.
- [Baker et al 1998] P. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, and R. Stevens, TAMBIS-Transparent Access to Multiple Bioinformatics Information Sources, *In 6th International Conference on Intelligent Systems for Molecular Biology,* AAAI Press, Montreal, Canada, 1998.
- [Balabanovic & Shoham 1995] M. Balabanovic and Y. Shoham, Learning Information Retrieval Agents: Experiments with Automated Web Browsing, *In On-line Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, Heterogeneous Environments*, 1995.
- [Baxevanis02] A.D. Baxevanis, The Molecular Biology Database Collection: 2002 Update, *Nucleic Acids Res.* 2002 30: 1-12
- [Benson et al 2002] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL., *Nucleic Acids Res* 2002, 30(1):17-20 GenBank.
- [Berners-Lee 2001] The Semantic Web, *Scientific American, May 2001*, Tim Berners-Lee, James Hendler and Ora Lassila.
- [BioNetbook] http://www.pasteur.fr/recherche/BNB/bnb-en.html
- [Birmingham et al 1994] W.P. Birmingham, C.O. Frost, A.J. Warner, and K. Willis, The University of Digital Library: This is not your father's library*, Proceedings of Digital Libraries*, 1994, 53-60.
- [Biswas & Lee 1994] G. Biswas and G. Lee. Knowledge Reorganization: A Rule Model Scheme for Efficient Reasoning, Proc. *Tenth IEEE Conference on AI for Applications (CAIA)*, San Antonio, TX, pp. 312-318, March 1994.
- [Biswas et al 1998] G. Biswas, J. Weinberg, and D. Fisher, ITERATE: A Conceptual Clustering Algorithm for Data Mining, I*EEE Trans. on Systems, Man, and Cybernetics*, Vol. 28, Pt. C, No.2, May, 1998.
- [Bolter 1991] J.D. Bolter, Writing space: *The Computer, Hypertext, and the History of Writing*, New Jersey: Lawrence Erlbaum Associates Publishers, 1991.
- [Bouziane & Hsu 1997] M. Bouziane and C. Hsu, A Rulebase Management System Using Conceptual Rule Modeling, *International Journal on Artificial Intelligence Tools*, 6(1), 1997, 37-61.
- [Bowman et al 1994] C.M. Bowman, P.B. Danzig, D.R. Hardy, U. Manber and M.F. Schwartz, The Harvest Information Discovery and Access System, *Proceedings of the Second International World-Wide Web Conference*, Chicago, Illinois, Oct 1994.
- [CACM 1994] CACM, Dexter hypertext issue, *Communications of the ACM*, 37: 2, February, 1994.
- [Chen & Sycara 1998] L. Chen and K. Sycara, WebMate: A Personal Agent for Browsing and Searching, *Proc. Of the 2nd International conference on Autonomous Agents*, 1998
- [Clark & Niblett 1989] P. Clark and T. Niblett, The CN2 Induction Algorithm,

Machine Learning, 3(1989), 261--283.

- [D-Lib] The Working Group on Digital Library Metrics. http://www.dlib.org/metrics/public.
- [Fensel et al 2003] *Spinning the Semantic Web Bringing the World Wide Web to Its Full Potential* Edited by Diter Fensel, James Hendler, Henry Lieberman, and Wolfgang Wahlster The MIT Press,
Cambridge, Massachusetts 2003.
- [Fikes & Kehler 1985] R. Fikes and T. Kehler, The Role of Frame-Based Representation in Reasoning, *Communications of the ACM*, 28(1985), 9: 904-920.
- [Frisse & Cousins 1992] M. Frisse and S. Cousins, Models for hypertext, *Journal of the American Society for Information Science*, 43: 2, 183--191, March, 1992.
- [Fuhr et al 2001] N. Fuhr, P. Hansen, M. Mabe, A. Micskik, and I. Solvberg, Digital libraries: A Generic Classification and Evaluation Scheme, *European conference on Digital Library*, 2001
- [Fujibuchi 1997] W. Fujibuchi, DBGET/LinkDB: A way of Solution to Integrate Diverged Biological Databases, *ICR Annual Report*, Vol. 4., 1997.
- [GDB] http://gdb.jst.go.jp/gdb/gdbDataModel.html
- [Galperin 2004]. M. Y. Galperin: The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res.* 2004 Jan 1;32 Database issue:D3-22
- [Gellersen et al 1997] H.-W. Gellersen et al, WebComposition: An Object-Oriented Support System for the Web Engineering Life Cycle, *WWW6 International Conference Poceedings 1997*,
http://www6.nttlabs.com/HyperNews/get/PAPER232.html.
- [GO] Gene Ontology Consortium. http://www.geneontology.org
- [Giarratano 1993] J. Giarratano, *CLIPS User's Guide (CLIPS Ver 6.0),* Lyndon B. Johnson Space Center, Information Systems Directorate, *Software Technology*, NASA, USA, 1993.
- [Google] http://www.google.com
- [Grϕnbæk & Trigg 1994] K. Grϕnbæk, and R. Trigg, Design Issues for a Dexter-Based Hypermedia System, *Communications of the ACM*, 37: 2, 40--49 February, 1994.
- [Gupta et al 2000] Knowledge-Based Integration of Neuroscience Data Sources, 12th Int. *Conference on Scientific and Statistical Database Management*, Berlin, Jul. 2000.
- [Halasz 1988] F.G. Halasz, Reflections on NoteCards: Seven issues for the next generation of hypermedia systems, *Communications of the ACM*, 31: 7, 836--852, July 1988.
- [Halasz & Schwartz 1994] F.G. Halasz and M. Schwartz, The Dexter Hypertext reference model, *Communications of the ACM*, 37: 2, 30--39, February, 1994.
- [Han & Fu 1999] J. Han and Y. Fu, Multiple-Level Association Rules, *IEEE Transactions on Knowledge and Data Engineering*, Volume 11, Number 5, October 1999.
- [ILOG 1998] ILOG, ILOG Rules, 1998. URL:
http://www.ilog.com/html/products/infrastructure /rules.htm.
- [Ingham et al 1995] D.B. Ingham et al, Bringing Object Oriented Technology to the Web, *WWW4 International Conference Proceedings*, 1995,
http://w3objects.ncl.ac.uk/pubs/bootw/.
- [Ingham et al 1997] D.B. Ingham et al, Supporting Highly Manageable Web Services, *WWW6 International Conference Proceedings, 1997*,
http://w3objects.ncl.ac.uk/pubs/shmws/.
- [ISMB 1998] Semantic Foundations for Molecular Biology Schemata, Controlled

Vocabularies and Ontologies, *Workshop at the Sixth International Conference on Intelligent Systems for Molecular biology*, 1998.

- [Joachims et al 1995] T. Joachims, T. Mitchell, D. Freitag, and R. Armstrong, WebWatcher: Machine Learning and Hypertext, GI Fachgruppentreffen Maschinelles Lernen, K. Morik and J. Herrmann (Eds.), University of Dortmund, Germany, August 1995.
- [Joachims et al 1997] T. Joachims, D. Freitag, and T. Mitchell, WebWatcher: A Tour Guide for the World Wide Web, *Proceedings of the 15th International Conference on Artificial Intelligence*, Nagoya, Japan, August 23-29, 1997, 770-775.
- [Kanehisa & Goto 2000] Kanehisa, M. and Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., 28, 27-30.
- [Karp, Paley, & Romero 2002] P. Karp, S. Paley, and P. Romero: The Pathway Tools Software, *Bioinformatics* 18:S225-32 2002.
- [Kilov 1994] H. Kilov, On Understanding Hypertext: Are Links Essential? *ACM SIGSOFT Software engineering notes*, 19: 1, January, 1994.
- [Koster 1994] M. Koster, ALIWEB - Archie-Like Indexing in the Web, *Proceedings of the First International World-Wide Web Conference*, Geneva Switzerland, May 1994.
- [Koster 1995a] M. Koster, Robots in the Web: threat or treat? *ConneXions*, Volume 9, No 4, April 1995.
- [Koster 1995b] M. Koster, Guidelines for Robot Writers, http://info.webcrawler.com/mak/projects/ robots/guidelines.html.
- [Koster 1995c] M. Koster, A Standard for Robot Exclusion, http://info.webcrawler.com/mak/ projects/robots/norobots.html.
- [Krulwich 1995] B.T. Krulwich, An Agent of Change, Andersen Consulting, http://bf.cstar.ac.com/bf/article1.html.
- [Landow 1992] G.P. Landow, *HYPERTEXT: The Convergence of Contemporary Critical Thought and Technology*, Baltimore: The John Hopkins University Press, 1992.
- [Lee 1994] Lee, G., Increasing Reliablity & Efficiency for Knowledge Based Systems, Ph.D. Dissertation, Vanderbilt University, 1994.
- [Lee 1999] Lee, G., Construction of Stable Clustering by Minimizing the Order Bias, *Korea information Processing Society,* Vol. 6, No. 6, 1999.
- [Leser et al 1998] U. Leser, R. Wagner, A. Grigoriev, H. Lehrach and H.R. Crollius, IXDB, An X Chromosome Integrated Database, *Nucleic Acids Research*, Vol. 26 No. 1, 1998.
- [Loke et al 1996] S. W. Loke, A. Davison and L. Sterling, CIFI: An Intelligent Agent for Citation, *Technical Report 96/4*, Department of Computer Science, The University of Melbourne, Parkville, Victoria 3052, Australia.
- [Ludascher et al 1998] Managing Semi-structured Data with Florid: A Deductive Object-Oriented Perspective, *Information Systems*, vol. 23, no. 8, pp 1-25, 1998, Elsevier Science Ltd.
- [Lycos] http://www.lycos.com
- [Markowitz et al 1999] OPM: Object-Protocol Model Data Management Tools. *Bioinformatics: Database and Systems*, S. Letovsky (ed.), Kluwer, Norwell, MA, pp. 187-199.
- [May & Lausen 2000] W. May and G. Lausen, Information Extraction from the Web, *Technical report no. 136*, Institut fur Informatik, Germany, Mar., 2000.
- [Michalski et al 1986] R.S. Michalski, I. Mozetic, J. Hong, and N. Lavrac, The Multi-Purpose Incremental Learning System AQ15 and Its Testing Application to Three

Medical Domains, *Proceedings of AAAI 1986*, 1986, 1041--1045.

- [Mitchell 1978] T.M. Mitchell, Version Spaces: An Approach to Concept Learning, Ph.D. Thesis, Stanford University, 1978.
- [Morris & Maes 2000] J. Morris and P. Maes, Sardine: An Agent-facilitated Airline Ticket Bidding System, Software Demos, *Proceedings of the Fourth International Conference on Autonomous Agents (Agents 2000),* Barcelona, Catalonia, Spain, June 3 - June 7, 2000.
- [Moss 1994] C. Moss, *Prolog++: The Power of Object-Oriented and Logic Programming*, Addison-Wesley, 1994.
- [Murphy et al 2004] Murphy JA, Barrantes-Reynolds R, Kocherlakota R, Bond JP, Greenblatt MS, The CDKN2A Mutation Database, *Human Mutation* in press 2004.
- [Neil et al 1999] N.W. Van Dyke, H. Lieberman, and P. Maes, Butterfly: A Conversation-Finding Agent for Internet Relay Chat, *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, January 1999, Redondo Beach, CA, 1999
- [Ngu & Wu 1998] D.S.W. Ngu and X. Wu, Interest Discovery for Incremental Web Exploration, *Proceedings of WebNet 98: The 1998 World Conference of the WWW, Internet and Intranet*, Orlando, Florida, USA, November 7-12, 1998.
- [Nguyen, Wu & Sajeev 1998] T.-L. Nguyen, X. Wu and S. Sajeev, Object-Oriented Modeling of Multimedia Documents, *Proceedings of the Seventh International World Wide Web Conference (WWW7)* (Brisbane, Queensland, Australia, 14 - 18 1998), published in *Computer Networks and ISDN Systems: The International Journal of Computer and Telecommunications Networking*, 30(1998): 578-582.
- [Nica & Rundensteiner 1996] Anisoara Nica and Elke A. Rundensteiner, Uniform Structured Document Handling using a Constraint-based Object Approach, In: *Digital Libraries: Research and Technology Advances* (Lecture Notes in Computer Science 1082), N.R. Adam, B.K. Bhargava, M. Halem, and Y. Yesha (Eds.), Springer-Verlag, Berlin, 1996, 27--34.
- [Oliver et al 2002] Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P, Hum Mutat 2002 19(6):607-14, *The IARC TP53 Database: New Online Mutation Analysis and Recommendations to Users.*
- [Olsson et al 2000] T. Olsson, A. Rasmusson, and S. Janson, Personalized Decentralized Communication. AAAI Spring Symposium Series 2000
- [OMIM] Online medlelian inheritance in Man, *NCBI*, http://www.ncbi.nlm.nih.gov/omim
- [Paepcke et al 1996] A. Paepcke, S.B. Cousins, H. Garcia-Molina, S.W. Hassan, S.P. Ketchpel, M. Roscheisen, and T. Winograd, Using Distributed Objects for Digital Library Interoperability, *Computer*, 29(1996), 5: 22-26.
- [Paepcke et al 1998] A. Paepcke, C.-C. K. Chang, H. Garcia-Molina, and T. Winograd, Interoperability for Digital Libraries Worldwide, *Communications of the ACM*, 41(1998), 4: 33-43.
- [Pazzani et al 1995] M. Pazzani, L. Nguyen and S. Mantik, Learning from Hot Lists and Cold Lists: Towards a WWW information filtering and seeking agent, *Proceedings of IEEE 1995 Intl. Conference on Tools with AI*, 1995.
- [Pazzani et al 1996] M. Pazzani, J. Muramatsu and D. Billsus, Syskill & Webert: Identifying interesting Web sites, AAAI Spring Symposium on Machine Learning in Information Access, Technical Papers, Stanford, March 25-27, 1996.
- [Pitkow & Bharat 1994] J. E. Pitkow and K. A. Bharat, WebViz: A Tool for WWW Access Log Analysis, *Proceedings of the First International World-Wide Web Conference*, Geneva Switzerland, May 1994.

- [Quinlan 1993] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
- [Rasmusson et al 1998] A. Rasmusson, T. Olsson, and P. Hansen, A Virtual Community Library: *SICS Digital Library Infrastructure Project, Second European Conference, ECDL'98*, Heraklion, Crete, Sept, 1998.
- [Riecken 1994] D. Riecken, Intelligent Agents, *Communication of the ACM*, Vol 37, No 7, July 1994.
- [Russo et al 2001] F. Russo, G. Pelts, E. Evans, R. Shank, and L. Grover, Genomic Knowledge Platform: An Object-Oriented Architecture Integrating Information Across Sequence, Expression, Genetics, and Other Biological Data Sources, *Object in Bio & Chem-Informatics 2001 (OiBC-2001),* Boston, MA, July. 2001
- [Safran et al 2001] M. Safran, S. Shen-Orr, I. Solomon, I. Peter, V. Chalifa-Caspi, and D. Lnacet, GenCards 3.0: An Object-Oriented Approach, *Object in Bio & Chem-Informatics 2001 (OiBC-2001)*, Boston, MA, July. 2001
- [Schatz et al 1996] B. Schatz, W.H. Mischo, T.W. Cole, J.B. Hardin, A.P. Bishop, and H. Chen, Federating Diverse Collections of Scientific Literature, *Computer*, 29 (1996), 5: 28-36.
- [Schulze-Kremer 1998] S. Schulze-Kremer, Ontologies for Molecular Biology, *Proc. Of the Third Pacific Symposium on Biocomputing*, Hawaii, AAAI Press, pp. 693-704.
- [Schwabe & Rossi 1995] D. Schwabe and G. Rossi, The Object-Oriented Hypermedia Design Model, *Communications of the ACM*, 38: 8, 45-46, August, 1995.
- [Scriver et al 2000] Scriver CR, Nowacki PM, Lehvaslaiho H, Hum Mutat 2000;15(1):13-15, Guidelines and Recommendations for Content, Structure, and Deployment of Mutation Databases: II. Journey in progress.
- [Shapiro 1987] A.D. Shapiro, Structured Induction in Expert Systems, Turing Institute Press in association with Addison-Wesley, 1987.
- [Shoham 1993] Y. Shoham, Agent-Oriented Programming, *Artificial Intelligence*, 60(1993), 51-92.
- [Stone et al 2004a] Jeffrey Stone, Xindong Wu, and Mark Greenblatt, A Semantic Network for Modeling Biological Knowledge in Multiple Databases, *Communications of the IIMA*, 4(2004), 4: 41-57.
- [Stone et al 2004b] Jeffrey Stone, Xindong Wu, and Mark Greenblatt, An Intelligent Digital Library System for Computational Biologists, *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB2004)*, Stanford, CA, August 16-19, 2004, 491-492.
- [Stotts & Furuta 1991] P.D. Stotts and R. Furuta, Hypertext 2000: Databases or Documents? *Electronic publishing*, 4: 2, 119--121, 1991.
- [Utgoff 1989] P.E. Utgoff, Incremental Induction of Decision Trees, *Machine Learning*, Vol 4, 1989, 161-186.
- [UMLS] National Library of Medicine's Unified Medical Language System. http://www.nlm.nih.gov/research/umls.
- [W3C] http://www.w3.org/RDF/Overview.html.
- Wactlar et al 1996]{cmu} H.D. Wactlar, T. Kanade, M.A. Smith, and S. Stevens, Intelligent Access to Digital Video: Informedia Project, *Computer*, 29(1996), 5: 46-52.
- [WebObjects 1997] WebObjects, http://www.next.com/WebObjects/.
- [Witten 2000] *Data Mining: Practical Machine Learning Tools with Java Implementations*, by Ian H. Witten and Eibe Frank, Morgan Kaufmann, San Francisco, 2000.
- [World Wide Web Worm] http://www.cs.colorado.edu/home/mcbryan/WWWW.html

- [Wu 1993] X. Wu, The HCV Induction Algorithm, *Proceedings of the 21st ACM Computer Science Conference*, S.C. Kwasny and J.F. Buck (Eds), ACM Press, 1993, 169-175.
- [Wu 1995] X. Wu, *Knowledge Acquisition from Databases*, Ablex Publishing Corp., USA, 1995.
- [Wu 1997] X. Wu, Interpretation of `No Match' and `Multiple Match' in Induction, *The Computer Journal*, 40(1997), 1: 50-57.
- [Wu 1998] X. Wu, Rule Induction with Extension Matrices, *Journal of the American Society for Information Science*, 49(1998), 5: 435-454.
- [Wu et al 1995] X. Wu, S. Ramakrishnan, and H. Schmidt, Knowledge Objects, *Informatica: An International Journal of Computing and Informatics*, 19: 4, 1995, 557--571.
- [Wu et al 1996] X. Wu, J. Krisar, and P. Mahlen, Noise Handling with Extension Matrices, *International Journal on Artificial Intelligence Tools*, 5 (1996), 1: 81-97.
- [Wu & Cai 2000] X. Wu and K. Cai, Knowledge Object Modeling, *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans,* 30(2000), 2: 96-107.
- [Yahoo] http://www.yahoo.com/
- [Yan et al 1996] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina and Umeshwar Dayal, From user access patterns to dynamic hypertext linking, *Proceeding of the Fifth International World Wide Web Conference*, Paris, France, May 1996.
- [Zhao, Quek & Wu 1998] M. Zhao, F. Quek and X. Wu, RIEVL: Recursive Induction Learning in Hand Gesture Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1998), 11: 1174-1185.

Appendix 1. Semantic Types

Bold items indicate additions to the standard UMLS Semantic Network.
Bold Italicized items indicate additions that are not part of the Gene Ontology.

- 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**
          - **Unlocalized**
        - Gene or Genome
    - Substance
      - Chemical
        - Chemical Viewed Functionally
          - Phamacologic Substance
            - Antibiotic
            - Clinical Drug
          - Biomedical Material

- [Entity] (continued)
  - [Physical Object] (continued)
    - [Substance] (continued)
      - [Chemical] (continued)
        - [Chemical Viewed Functionally] (continued)
          - Biologically Active Substance
            - Neuroreactive 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
  - 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 of 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**
        - **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**
          - **Gycosaminoglycan 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**
- **Transporter Activity**
- **Triplet Codon-AA Adaptor Activity**

- Pathologic Function
  - Disease or Syndrome
    - Mental or Behavioral Dysfunction
    - Neoplastic Process
  - Cell or Molecular Dysfunction
  - Experimental Model of Disease
- Injury or Poisoning

Appendix 2. Semantic Relationships

Bold items indicate additions to the standard UMLS Semantic Network.
Bold Italicized items indicate additions that are not part of the Gene Ontology.

- 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***
            - Performs
                - Carries out
                - Exhibits
                - Practices
            - Occurs in
                - Process of
            - Uses
            - Manifestation of
            - Indicates
            - Result of
        - Temporally related to
            - Co occurs with
            - Precedes

- [Is a] (continued)
  - [Associated with] (continued)
    - Conceptually related to
      - Evaluation of
      - Degree of
      - Analyzes
        - Assesses effect of
        - Measures
        - Diagnoses
        - Property of
        - Derivative of
        - Developmental form of
        - Method of
        - Conceptual part of
        - Issue in
    - ***Similarity related to***
      - ***Functionally simular to***
      - ***Physically similar to***
        - ***1D Structure related to***
        - ***2D Structure related to***
        - ***3D Structure related to***
        - ***4D Structure related to***