

# An Intelligent Digital Library System for Biologists

Jeffrey Stone  
Dept. of Computer Science,  
University of Vermont  
jestone@zoo.uvm.edu

Xindong Wu  
Dept. of Computer Science,  
University of Vermont  
xwu@emba.uvm.edu

Marc Greenblatt  
Department of Medicine,  
University of Vermont  
Marc.Greenblatt@vtmednet.org

## 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. Our digital library system is centralized with Web links to publicly accessible data repositories. Our digital library is based on a framework used for conventional libraries and an object-oriented paradigm, and will provide personalized user-centered services based on the user's areas of interests and preferences. To make personalized service possible, a "user profile" that represents the preferences of an individual user is constructed based upon the user's 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 with minimal user effort.*

## 1. Introduction

Recent advances in the fields of computational biology, cloning and genetics have resulted in vast amounts of data, which are providing an unprecedented volume of knowledge to researchers and medical personnel. This information will be critical for the understanding of biological structure and function and has allowed for the development of new treatment approaches for disease such as gene therapy and pharmacogenetics. However, the amount of data that the researcher must digest on a daily basis has become unmanageable. Furthermore, the view of biological phenomena as being composed of a number of sub disciplines (e.g., structural biology, genomics, proteomics, and biochemistry) has served to further complicate the issue. To obtain a coherent picture of biological phenomena at the molecular, cellular, and organism levels, one must

both look at all of these attributes and at 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.

To aid the researcher in this task of information retrieval and organization, we are developing a sophisticated digital library system that utilizes data mining techniques and user profiling to recommend items to the user. Our digital library system is centralized with Web links to publicly accessible data repositories. Based on the framework of a conventional library, our system will also provide user-centered services based on a user's past activities and preferences.

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. The semantic network will impart a knowledge structure through which the system can "reason" and draw conclusions about biological data objects and will provide a federated view of the many disparate databases of interest to biologists. In the development of our 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.

## 2. Library System Design

Our approach begins from the centralized, structured view of a conventional library, and seeks to provide access to the digital library through electronic means including the Internet, while maintaining the advantages of decentralization, rapid evolution and flexibility of the Web. 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 the user’s 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.

### 3. Semantic network based dictionary

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. Our semantic network strives to provide a categorization of biological concepts and relationships among these concepts. The semantic network will impart a knowledge structure through which 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 [1]. 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 [2].

### 4. Generation of rules

The rules that are used for the recommendation of items in our system are generated with the popular open source Weka data mining package [2]. 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. This program will 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 after generation.

### 5. Conclusions

Through the combination of data mining, user profiles and a semantic network, our system can aid researchers in obtaining, organizing and managing biological data. Through the use of recommendation agents, our system will help the novice researcher discover new topics and items to look at, and the expert to quickly view new items that pertain to their research area.

### 6. References

- [1] NLM Unified Medical Language System, National Library of Medicine, <http://www.nlm.nih.gov/research/umls>.
- [2] Gene Ontology Consortium. <http://www.geneontology.org>
- [3] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.