Project: High Throughput Data Retrieval from Biological Databases

Students: Hyesun Cha, Lam Tran, Nan Wu
Mentor: Joseph Brown
Sponsor: Pacific Northwest National Laboratory
Instructor: Behrooz Shirazi

Course: CptS 421

Date Summited: February 12, 2013

Review of Literature

Project Objective

This project will develop a framework in C# (.net 4.0+) that provides connectivity to NCBI and other open data repositories using web-based technologies, Entrez Programming Untilities, to high throughput data retrieval. We will also develop several GUI elements using WPF to display these results, and provide researchers powerful tools for understanding which proteins have or have not been studied for particular diseases, such as HIV.

Project Overview

The National Center for Biotechnology Information (NCBI) is one of many data repositories that contain a plethora of biological data, articles, and links to scientific research. While full web-sites provide access to this data, their interfaces are often limited to single query searches and painstakingly manual.

Usually, people name protein by their functions, but some protein have more than one function. This means a single protein might have various names, makes searching more complex. One example of complex searching is MAVS also known as VISA, CARDIF, and IPS-1. To access all articles related to MAVS, researchers have to perform four individual searches. However, our SQLite database will combine the search result and reduce overlapping articles.

As shown in Figure 1, the implementation of this project can be divided into three main parts: networking, database and GUI. The program first will retrieve data in XML format from open data repository website, such as NCBI. After this process, we will parse data into SQLite database. After getting input from user, GUI will interact with database and return a hit count summary table which contains count number of related articles. To further access to specific article information, user need to click on article numbers and GUI will provide articles details.

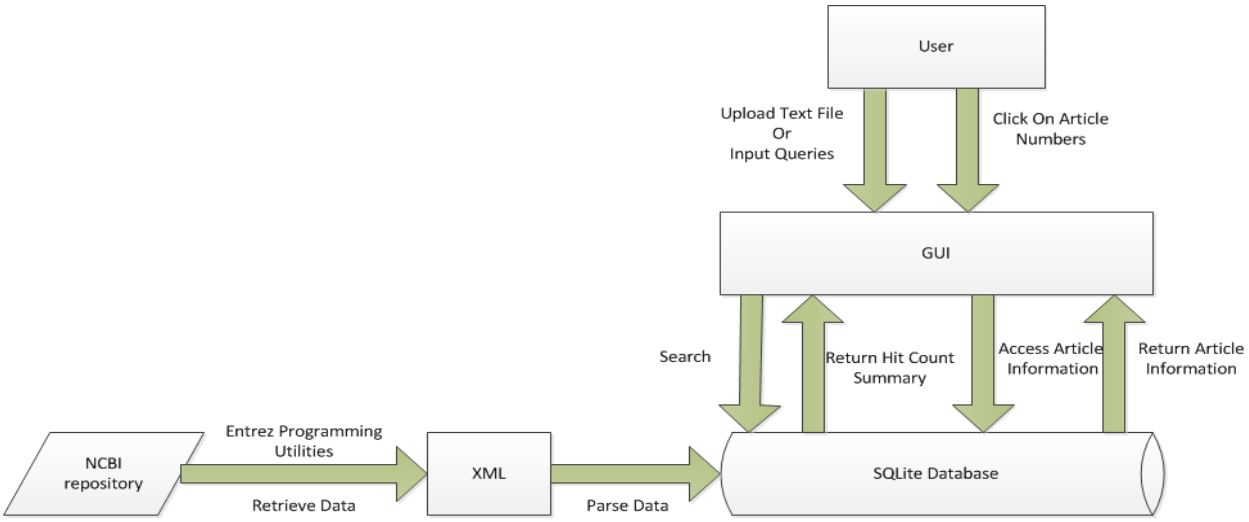

Figure 1 Project Flowchart Diagram

To search articles, we construct URLs based on what user input keywords or load a text file on GUI, and connect the National Center for Biotechnology Information (NCBI) by the URLs. To retrieve data, we use Entrez Programming Utilities, which are a set of eight server-side programs that provide a stable interface into the Entrez query and database system at the NCBI. The E-utilities use the URLs syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. The E-utilities are therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature. After retrieval data from the NCBI, we store it into two types of XML format. First type will include the number of articles and a list of all articles which are found by the URLs, and the other type will be more detailed about a specific article which is a part of a list.

From NCBI, a huge number of articles in Pubmed database will be organized and stored in XML format. Each article typically has a title, a summary, author, date, etc. All of information of an article will be stored in hierarchical structure of XML file. The XML format has a schema just like databases, although as XML files are hierarchical in nature, the schema takes on a very different form. The program must be able to use the XML file as the input file to make database.

To build the database from XML file, we will use C# to parse the XML file, and then put all information in DataTables within a DataSet. These DataTables will be store as table in the SQLite database. It will provide the tool to make the data structure for database. The data structure will be a group of tables that include all information take from XML file, using PMID from article as the primary key for the relationship between each table (Figure 2).

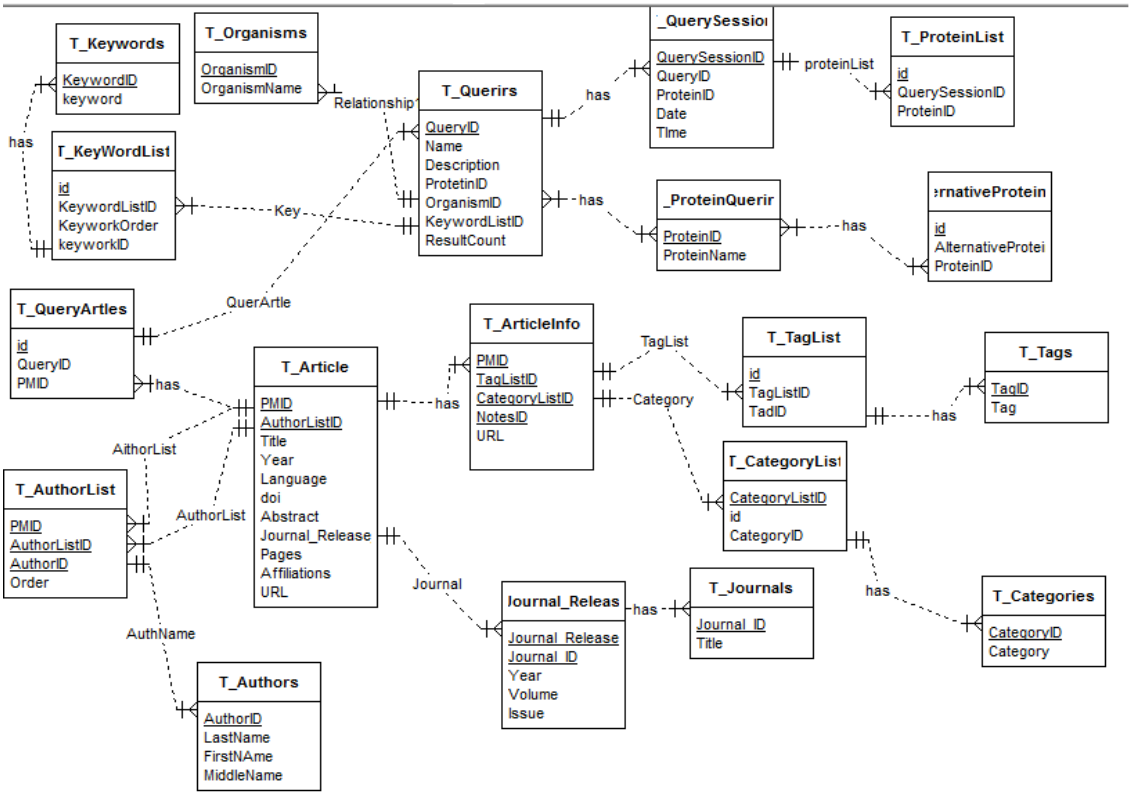

Figure 2 Database Relationship Diagram

Figure 3-1 to Figure 3-4 provides a general GUI view. GUI will be implemented by WPF and WVVM. To make GUI user friendly, we will provide two ways to get user input. User can either upload an existing text file containing protein names, organisms and keywords, or user can directly input in textboxes. Also, user can save workspace for later access. After clicking View Articles button, database will pass count numbers of articles that relate to inputs and generate a hit count summary table. User can click on data in hit count summary table and get access to article information. The number of articles will show on GUI. We will first provide user of lists of articles. User can filter articles by certain criteria, search by keywords or clear article lists. After user picks certain article, the tool will bring up specific article abstraction, which contains article name, journal, abstraction, year etc.

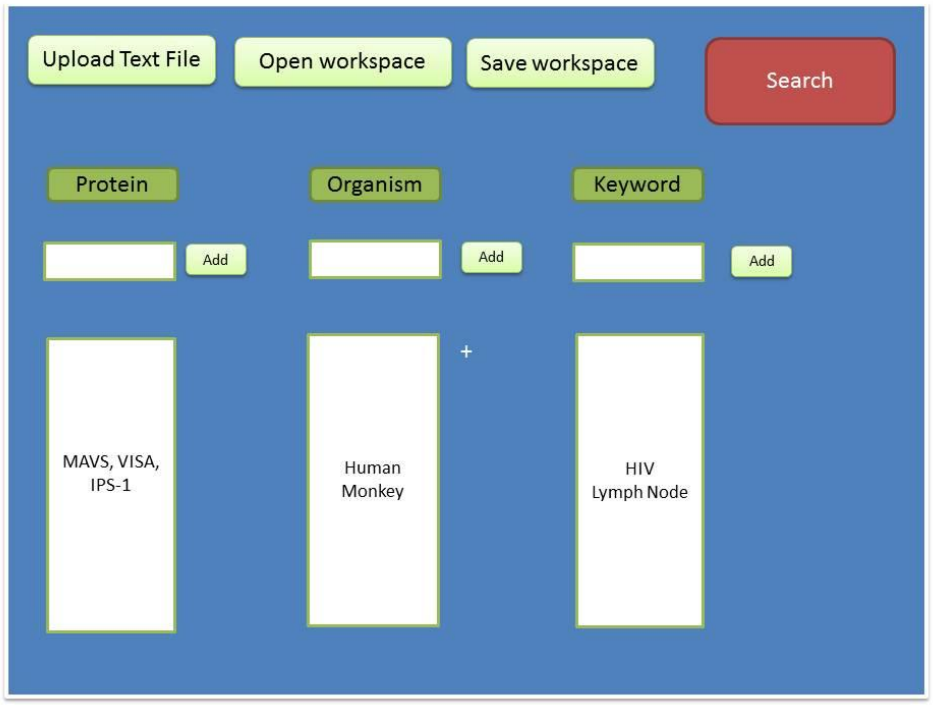

Figure 3-1 GUI View 1

Figure 3-2 GUI View 2

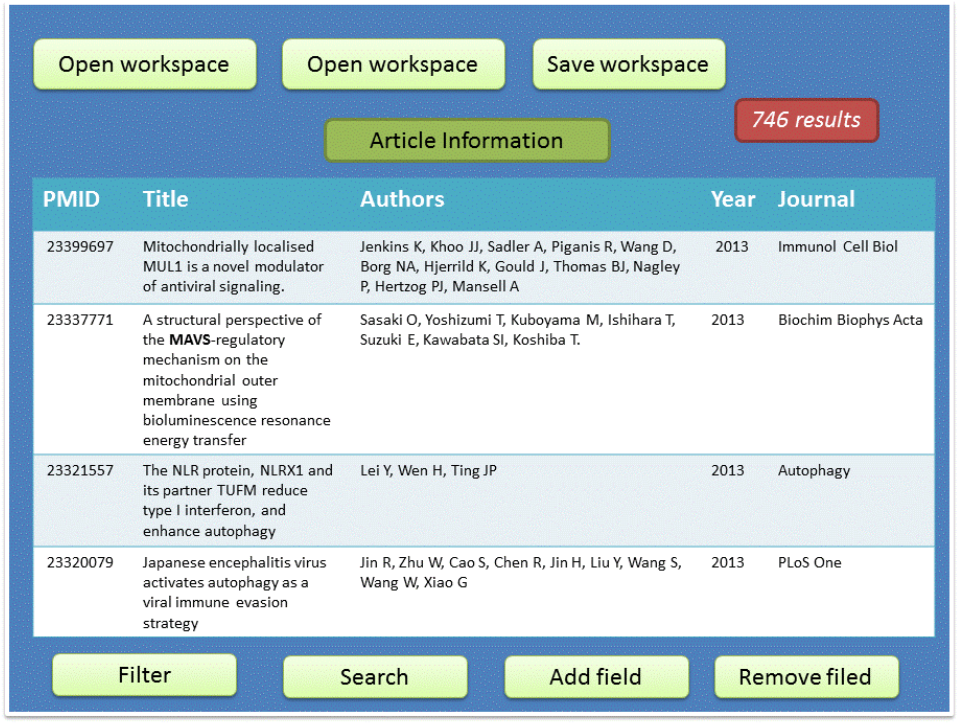

Figure 3-3 GUI View 3

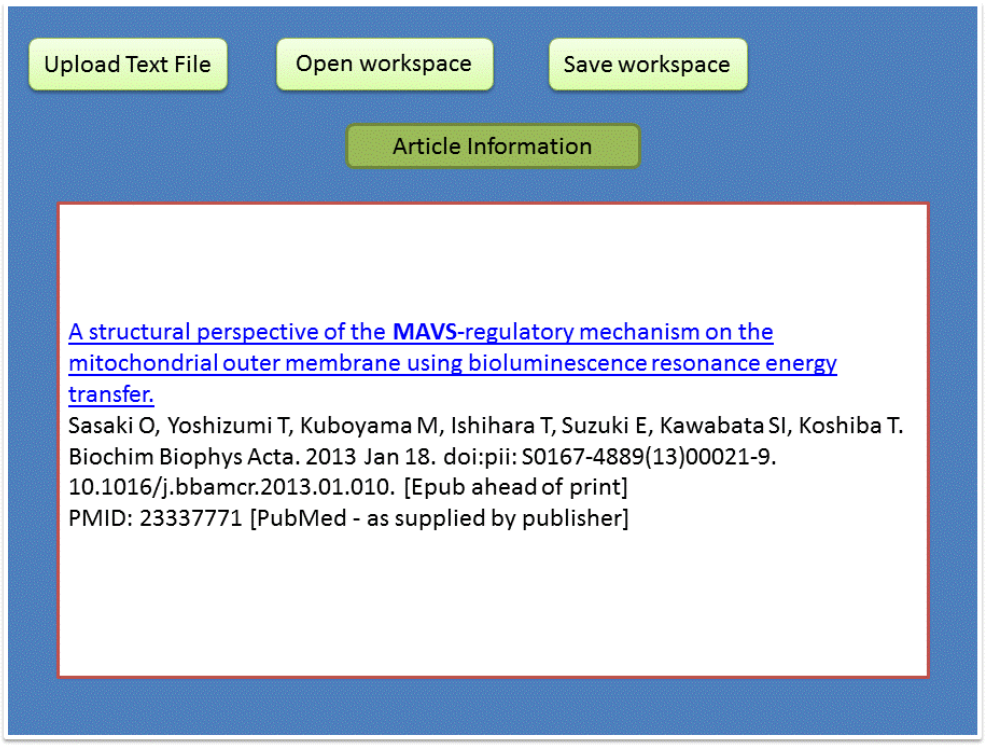

Figure 3-4 GUI View 4

Client Identification and Preferences

This project is designed to help researchers understand which protein have or have not been studied for certain disease. It will provide researchers with a more powerful and convenient way to search protein studies. After successfully building connection with NCBI repository, we will generalize other open data repositories and access to those repositories, such as EBI and JPDB. This framework will be developed to be used with other tools within software development team and will provide useful links into rich data repositories.

Stakeholder Identification and Considerations

We assume the stakeholders of this project are experienced researchers. With this project, scientist and researchers will benefit from user friendly interface that eliminates painstakingly manual and complex website search queries.

Objectives

- Utilize C# (.net 4.0) and MS Visual Studio 2010 as project implementation platform

- Retrieve data from NCBI website
    - Utilize Entrez Programming Utilities as data retrieving tool
    - Follow PubMed query style
    - Construct appropriate URLs based on input data
    - Store retrieved data into XML format
    - Parse XML files to store in SQLite database
    - Consider reaction time and correctness of retrieving data
- Utilize database platform by SQLite database
    - Create database from XML format
    - Design database structures to adapt the organization of information
    - Improve database for searching time
- Implement multiple layers, user friendly Graphical User Interface
    - Utilize WPF/MVVM to provide visualization
    - Design interface accepting both text file format and textbox input format
    - Clearly display data in organized table format
    - Improve visualization performance and reduce searching time compared to manual query search on NCBI website.
    - Provide intuitive way for users to manipulate data and understand results
- Distinguish differences between getting access to articles by local database and by directly link to website
    - Take advantage of most efficient method to manipulate data