An Effective System for Managing Biological Data

Qian Li
College of Computer and Control Engineering
College of Software
Nankai University, Tianjin, China
Email: liqian515@mail.nankai.edu.cn

Zhenglu Yang
College of Computer and Control Engineering
College of Software
Nankai University, Tianjin, China
Email: yangzl@nankai.edu.cn

Wei Cao
Department of Biotechnology
Graduate School of Agricultural and Life Sciences
The University of Tokyo, Japan
Email: davecao@bi.a.u-tokyo.ac.jp

Kentaro Shimizu
Department of Biotechnology
Graduate School of Agricultural and Life Sciences
The University of Tokyo, Japan
Email: shimizu@bi.a.u-tokyo.ac.jp

***Abstract*—Nowadays, life scientists grapple with a problem how to fast and easily access/obtain high quality data, especially for a specific research area, from a large amount of biological data deposited in public databases. In this work, we developed an effective system for managing biological data which are a class of functionally important membrane protein; they are hard to collected from the existing databases for slow progress of protein annotations, transitive annotation problem as well as low sequence similarity among them. Our preliminary system was designed with a user-friendly web interface and provides: 1) keywords retrieval against the annotation information of protein sequences, 2) recommendation of the related publications to help researchers conduct effective comparisons of experimental results with convenience, and 3) sequence alignment service (BLAST-based by NCBI blast+ and Hidden Markov Model-based by HMMER3.0). We had conducted a statistical analysis and showed it to the researchers in a visual way.**

1. Introduction

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis [1]. Biological database design, development, and long-term management is a core area of the discipline of bioinformatics [2]. While providing the data storage, the biological database provides a more convenient analysis and calculation tool for researchers, which promotes the development of biological science and the understanding of life.

There are many biological databases in the world, such as NCBI, Swiss-Prot and so on. These databases store the existing massive sequences and some of these public databases can provide users with data free of charge. But unfortunately, slow progress of protein annotations, transitive annotation problem, and low sequence similarity, especially for a specific research area for biological fields, make it hard for researchers to navigate these databases and obtain the necessary data rapidly. Here we have curated the functionally important proteins from three species and developed a specialized database of the proteins to provide reliable information and easily access to biologists or bioinformaticians.

2. Data and analysis

Each protein deposited in our system includes basic information, such as protein name, amino acid sequence and subcellular localization, and also manually curated literature information. To make a high-quality dataset, we manually read each paper associated with the specific biological topic, and extract the information of the protein which are demonstrated experimentally in the contexts of the paper. To our surprise, the proteins show low sequence similarity when we used the sequence alignment program, Blast++ v(2.3.0+) [3], to find homologous sequence against the Swiss-Prot database (reviewed entries). We could not detect homolgous sequences even with parameters of the higher identity and e-value (identity of 80% and e-value of 10^{-5}).

3. Function

Our system is a SQL-backend J2EE application which not only stores protein basic information and manually curated literature information but also provides analysis tools for researchers. The following will introduce the architecture of the database and describe specific functions of our system.

1) Our system adopted a three-layer data management model which consists of a client layer, a server layer, and a database layer. In the client layer, we implemented dynamic

HTML pages with javascripts and css to enhance interactivity. The server layer handles the requests of the client layer where we used the JSP/Servlet technology to dynamically process the requests and construct the responses, and EJB for the business logic. In the database layer we used the JDBC technology to handle a database query from the server layer and perform the transactions with the SQL-backend database. This application we implemented was deployed on the Apache Tomcat application server. For basic security practices, the system keeps track of users' accesses including the maximum requests, the visitors' IP, the visiting time, the browser information, and so on. We aimed at designing a simple and intuitive web interface to provide better accessibility, optimal viewing experience and easy navigation for the user; the system integrated the most popular responsive framework, i.e., bootstrap, into the front end. Through detecting all sorts of electronic devices, desktop, laptop or mobile, the system will make the page automatically adjust to different devices.

2) We adopted a broadly used software package Lucene to support the keywords/terms searching against the database. Run a program to create an index for all the content to be used for queries which includes the protein name, the subcellular localization, the amino acid sequence, the amino acid sequence description, the amino acid sequence feature, the manually curated literature title, the manually curated literature author, and the manually curated literature description. Therefore, the researchers can easily retrieve the related proteins through this function.

3) Our system supports homology-based sequence search through the Blast program and the HMMER program at the backend. Blast uses a heuristic algorithm with faster speed and HMMER is based on the hidden Markov model which has a higher sensitivity. Researchers can choose a suitable tool according to their own needs.

4) The system simultaneously provide related literature recommendations when researchers use a keyword search function. From the PubMed database of NCBI, we collected information related to the proteins, mainly including the title, content and links, as the reservoir of related literature recommendations. First, we applied the TF-IDF algorithm to extract keywords from the content of a paper and obtain the first set of 0-3 keywords for each paper; second, we used Part-of-Speech tagging to extract nouns from the title to get the second set of keywords because the title represents the theme of a paper to a great extent; and finally, we obtained the resultant keywords by taking into account the keywords given by the author of a paper and the aforementioned two sets of the keywords. After the acquisition of the keywords of each paper, we calculate the similarity between keywords of each paper and search words, and then get final recommended papers according to the similarity ranking.

5) A statistical analysis of frequencies of several standard amino acids and functional domains analysis of all proteins deposited in the database are shown on the website in an interactive mode. The frequencies of amino acids of proteins are shown as a histogram so that the user can visually understand the distribution of amino acids in proteins. We use Nightingale rose to show frequencies of functional domains of proteins that users can clearly see the important functional domains; The notes of functional domains facilitate the further understanding of the domains. All charts are drawn with the help of ECharts, which is a JavaScript framework providing intuitive, vivid, interactive, and highly customizable data visualization.

4. Results and discussion

In order to give users a better experience, we tested the function of the website. In the process of building an index, it costed an average of less than 1 second. We randomly generated 200 key words to carry out the search and the average retrieval time was less than 500ms. This speed can meet the needs of users. We randomly selected 100 protein sequences to do time and sensitivity tests using HMMER and Blast. Blast spent less time than that consumed by the HMMER approach on average. The number of HMMER retrieval results was always greater than or equal to Blast. In the literature recommendation test, we used two evaluation metrics, i.e., Cosine and Jaccard, to find the similarity between the key words of literature and the search keywords. Jaccard was superior to Cosine in both the time and the accuracy.

5. Conclusion

As a proprietary database of biological data, our preliminary system provides not only data but also user-friendly services to researchers. High quality proteins are curated and deposited in our system. Full text search, homology-based sequence search, literature recommendation and data visualization are helpful for researchers to study proteins. All the functions perform satisfactorily. Next, we will continue to expand the database and optimize the algorithm and framework.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61070089 and No. 11431006, the Ministry of Education of Humanities and Social Science Project under Grant No. 16YJC790123, and the Research Fund for International Young Scientists under Grant No. 61650110510, and partially supported by Platform for Drug Discovery, Informatics, and Structural Life Science from Japan Agency for Medical Research and Development (AMED).

References

[1] Attwood, K. T., Gisel, Eriksson, and E. Bongcam-Rudloff, "Concepts, historical milestones and the central place of bioinformatics in modern biology: A european perspective," *Bioinformatics*, 2011.

[2] P. Bourne, "Will a biological database be different from a biological journal?" *Plos Computational Biology*, vol. 1, no. 3, pp. 179–81, 2005.

[3] NCBI, "Blast+ executables," 2015.