Design and Implementation of Query Builder and Asynchronous Submission System for NCBI Entrez Databases

Abo Yesayan
National Polytechnic University of Armenia
Yerevan, Armenia
e-mail: abo.yesayan@gmail.com

Robert Hakobyan
National Polytechnic University of Armenia
Yerevan, Armenia
e-mail: rob.hakobyan@gmail.com

ABSTRACT
This paper discusses the issues of retrieving and further processing of the genetic data from NCBI's (National Center for Biotechnology Information) Entrez search engine. The requirements of use of the NCBI server, the rules of formulating the requests and the related difficulties are discussed within the scope of present work. In addition, the frequently used formats of biological data and issues of data retrieval in discussed formats are also considered. In order to deal with issues mentioned above a system has been developed, which provides features like query building, its asynchronous submition and retrieved data processing. The system enables the comfortable design of complex queries through a graphical interface. The system allows to submit defined requests and receive corresponding data asynchronously. The developed system also allows executing code snippets for downloading data in desired modified format. The system enables efficient management and manipulation of the genetic data. Particularly, it automates the process of collecting large datasets.

Keywords
Genetic data, data asynchronous collection, NCBI, E-utilities, query builder.

1. INTRODUCTION
NCBI is one of the biggest open-access repositories for ~~the~~ biological data. It includes databases for DNA, RNA, proteins, scientific literature and for many other types of biomedical information [1][2]. It provides the user with numerous tools for accessing the resources in database. Nevertheless, effective and customizable retrieval of large and complex datasets from the NCBI databases, namely from GenBank, is still an open issue, despite of its great importance for many biological disciplines, including evolutionary, population and medical genetics [3,4]. One possible way to overcome this problem is to study NCBI API and develop custom scripts for retrieving desired data. We proposed a method and implemented a system which uses NCBI's API and allows collecting large datasets by executing complex queries containing sequential range conditions. For example, system parses AA12345-AA23456 range pattern into list of AA12345, AA12346, ... , AA23456 accession numbers and allows to get the whole search result as one complete dataset.

2. AIMS AND OBJECTIVES
Entrez Programming Utilities (E-utilities) is a public API to the NCBI Entrez system and allows accessing all Entrez databases. Entrez system represents a set of 8 services that handles the search from 38 databases of the NCBI. The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search and retrieve the requested data. Functionality of the Entrez and the methods of formulating queries are discussed in details in [5].

In general, receiving data from Entrez system is performed in 2 steps:
1. At first step, the request, including conditions and the title of the database in which the search should be executed are sent. In response, the server returns XML document with the number of results found, the list of unique identification numbers for each corresponding result, special keys (QueryKey and WebEnv) and other properties of the requested data.
2. At second step, the special keys and the list of identifiers from the first step are resent to the server for retrieving full data about each entry. The full response data contains the complete or partial genome, its mutations, some references and other related information.

All NCBI resources are free and publicly available, although partial restrictions applied to the use of its services, aimed to avoid server overloads. In the case of Entrez, the limitation is maximum 3 requests in a second. In the case of large sessions (more than 100 requests) requesting party is limited to one request during one second within the timeframe of 21:00 to 09:00 (EST) or on weekends. Violation of above described rules will result in blocking the requesting IP and the NCBI services will no longer be available for use. Taking into account that requests can be sent from the local network, using the same real IP, the possibility of violating the rules mentioned above is high. Thus, keeping up with the NCBI rules is also a highly considerable issue.

After the collection of the biological data, one of the issues of the further processing is to convert the data to necessary formats. Most of the bioinformatics software [6] use custom formats as input data, therefore, it is remarkably essential to automate the process of data conversion to an appropriate format.

Considering the requirements of the Entrez system, restrictions on their usage and importance of converting the data into the desired format, it was decided to create a system to solve the following problems:

- Building and executing Entrez queries through a graphical interface
- Ensure the compliance of the queries with the NCBI requirements
- Allow users to modify the data before downloading if needed using executable code snippets, edit the text to have the data in necessary format

3. IMPLEMENTATION & WORKFLOW

In order to solve the problems described above, a system is designed and developed, allowing to construct and submit a desired query using an easy to use graphical interface. After submission, the system converts the request in order to comply with the Entrez accepted format [5] of the query, puts it into the queue in order to execute when appropriate. The requests in the queue are submitted asynchronously to the external Entrez API, keeping all the requirements of using the NCBI system. After the response from Entrez is received, they are available for fetching. The system allows downloading data in frequently used formats such as FASTA, PHYLIP, NEXUS, etc. One of the main features of the developed system is the possibility of Python code snippet execution. Moreover, with applied snippet, it is possible to convert data to the appropriate format before downloading. The basic structure of the developed system presented in fig. 1.

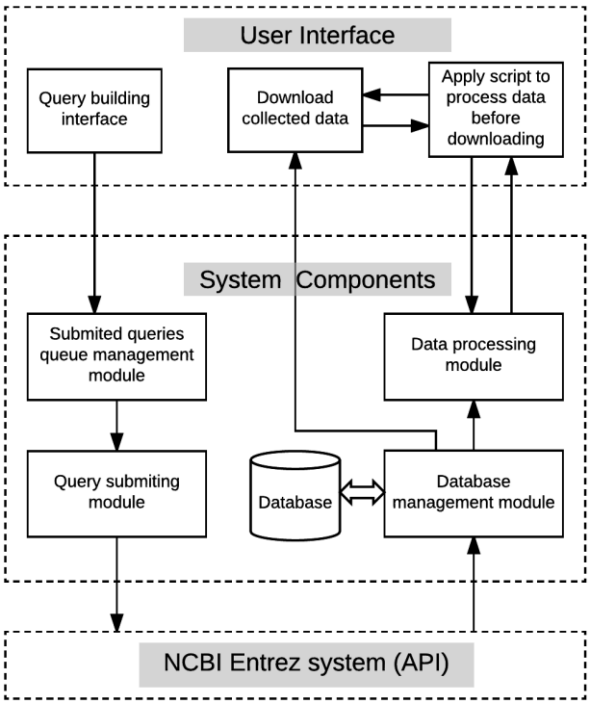

Fig. 1. Query builder and asynchronous submission system structure

For the system's software implementation Python programming language is used. The operation of the system in the web environment is supported by NGINX server [7] in compliance with the uWSGI interface [8], for data storage is used PostgreSQL [9] database management system. As an application containerization platform, Docker is used as a FOSS solution, which provides the ability to pack, run and deploy an application in a loosely isolated environment called a container [10].

4. CONCLUSION

In our study, Entrez system for the searching the NCBI's biological databases was analyzed, as well as the requirements and restrictions of its usage. To automatize and increase the effectiveness of large datasets retrieval for NCBI, we have designed and developed a system, which is using the capacity of Entrez search engine and provides a graphical interface for the asynchronous collection of data. Moreover, our system ensures the keeping all of the NCBI's terms of use while performing requests to Entrez. Another feature of the system is that it also allows applying Python code snippets to convert data to the appropriate format before downloading.

REFERENCES

[1] David L. Wheeler, Colombe Chappey, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, "Database resources of the National Center for Biotechnology Information" [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383943], *Oxford University, Nucleic Acids Research*, 2015.

[2] Benson DA1, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW, "GeneBank" [https:// www.ncbi.nlm.nih.gov/pubmed/23193287]. *Nucleic Acids Research 37*, 2013.

[3] Ewens W. J. "Mathematical Population Genetics (2nd Edition)", *Springer-Verlag, New York, ISBN 0-387- 20191-2*, 2004.

[4] Ca1valli-Sforza, L. Luca, and Marcus W. Feldman, "The application of molecular genetic approaches to the study of human evolution", *Nature Genetics 33*. 266 -275, 2003.

[5] Eric Sayers "A General Introduction to the E-utilities" [https://www.ncbi.nlm.nih.gov/books/NBK25497/] *Bethesda (MD): National Center for Biotechnology Information (US), 2010*.

[6] "List of Bioinformatics Software Tools for Next Generation Sequencing" [https://goo.gl/1q7hWQ]

[7] Rahul Sharma, "NGINX High Perfomance", 2015.

[8] "uWSGI Documentation Release 2.0" [https:// media.readthedocs.org/pdf/uwsgi-docs/latest/uwsgi-docs.pdf] [https://uwsgi-docs.readthedocs.io]

[9] Gregory Smith "PostgreSQL 9.0 High Performance", 2010.

[10] Adrian Mouat, "Using Docker. Developing and Deploying Software with Containers", *O'Reilly Media*, 2015.