RESEARCH ARTICLE OPEN ACCESS

A Secure Searching Technique of Nucleotide Sequence using BLAST Tool

Ipsita Saha [1], Joy Dewanjee [2], Miriam Ghosh [3]
Nazneen Hasan [4], Madhu Anand [5]
Department of Computer Science and Engineering
Gurunanak Institute of Technology
Sodepur, Panihati, Kolkata
West Bengal - India

ABSTRACT
The Basic Local Alignment Search Tool (BLAST) searches a particular database for the biological information depending on the input nucleotide sequence provided by the user. Now, the BLAST tool is available online for analysis. This paper deals with a faster approach to nucleotide sequence analysis with both online and offline compatibility by providing a personalized database for every user which fetches data remotely from the PDB and can be used for further analysis. While the BLAST only works with internet access and searches the database every time, this approach simply stores the information which is fetched by BLAST and all references regarding it from the online PDB database and keeps the track of the information securely. Now the users can concentrate on their analytical work, rather than accessing the online database every time which is saving a lot of time. The data which is fetched from the remote server is stored in an encrypted format in the offline database, allowing a complete security measure.

Keywords:- BLAST, Protein sequence, FASTA sequence, Protein Data Bank (PDB), Personalized Database.

I. INTRODUCTION

An organism physiology is vastly determined by its nucleotide sequence. It is a succession of letters that showed order of nucleotides. This biological information which direct the function of living organism by producing specific proteins, are digitally encoded into crystallographic databases like Protein Data Bank(PDB), Gene Bank, SWISS-PORT, etc. for studying and analytical purposes. Several algorithms like BLAST, FASTA, etc. have been developed to access these databases and display the nucleotide structure along with vital information.In the field of bioinformatics, BLAST is one of the most famous algorithm for DNA sequence searching. It's used for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. The algorithm takes input from the user as FASTA sequence and then passes the query to its server. The server then tactically searches the database in order to find match in the protein sequences and returns the result in plain html format. The result includes a detailed description of the every match in the input sequence and provides score value for determination of a perfect match along with the detailed information regarding the protein sequence fetched from the database.

II. RELATED WORK

In 1990, the BLAST algorithm was first published in the original paper by Altschul, et al .The computer program that implements the algorithm were developed by Stephen Altschul, Warren Gish, and David Lipman at the U.S. National Center for Biotechnology Information (NCBI), Webb Miller at the Pennsylvania State University, and Gene Myers at the University of Arizona. It is available on the web on the NCBI website. Although many alterations of the algorithm were done in order to improve its different shortcomings and thus making the algorithm more optimize. Alternative implementations include AB-BLAST (formerly known as WU-BLAST), FSA-BLAST (last updated in 2006), and ScalaBLAST. Later many algorithms were proposed including UBLAST, USEARCH which concentrate on a particular format of alignment search.

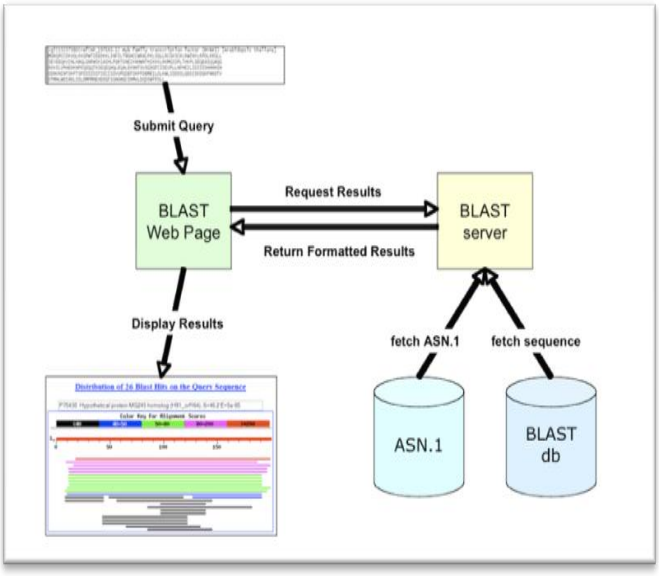


Fig. 1. Mechanism of BLAST tool

III. PROPOSED WORK

A. Purpose of the proposed approach

Purpose of this approach is mainly to include a client-side software along with an offline database which will cache the data recently viewed in the BLAST web page and cross reference the data with Protein Data Bank (PDB) to provide the users a detailed idea about their queries.

B. Proposed Workflow

The entire methodology of the software is described into two sub-modules. Each sub-module provides the different aspects of the entire functioning of the software. The workflow shows the entire mechanism of how the software works and stores the data through a simple flowchart implementation.

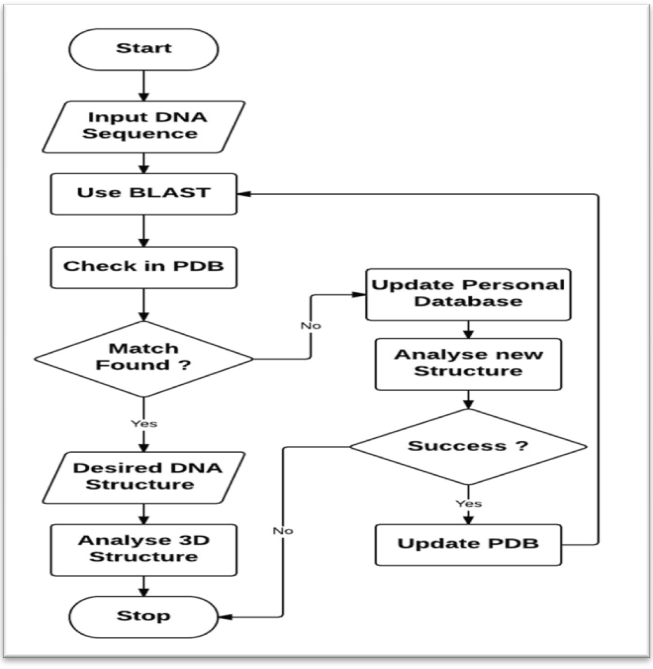
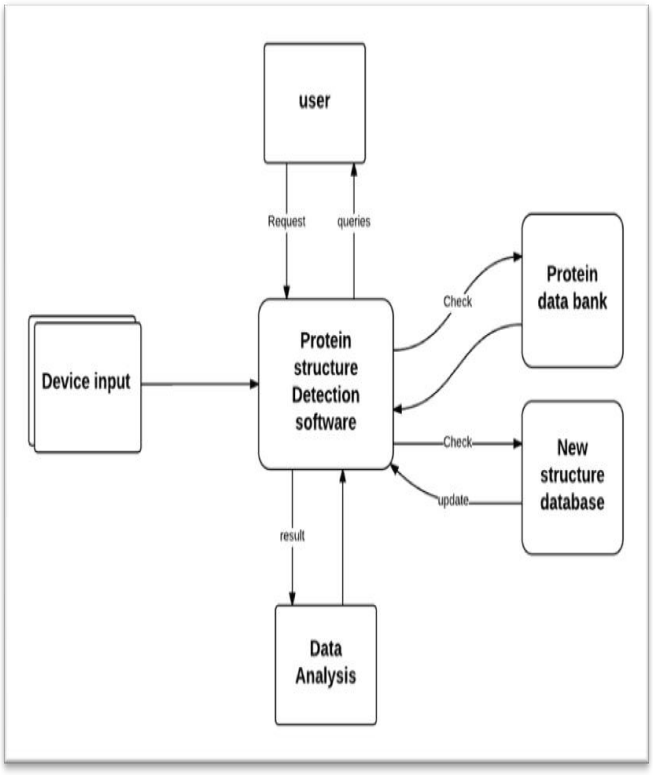


Fig. 2. Explanation of the application workflow

C. Proposed Dataflow

The dataflow gives a brief idea of how the how the data is



accessed and retrieved in the application by a user.

Fig. 3. Data Flow Diagram

D. Encryption-Decryption Technique

The data which is stored in the offline database is encrypted by using MD5 Hash algorithm so that the data remains secured in the offline database. The data is encrypted at the server side to avoid loss of data while transaction.

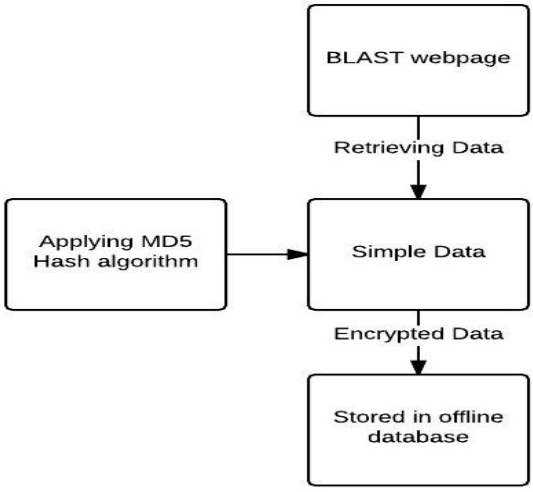


Fig. 4. Encryption of data for offline storage

E. Application Prototype

The idea of the software is implemented in this simple prototype which showcases how the protein sequence can be viewed at the front-end application connected to the offline database.

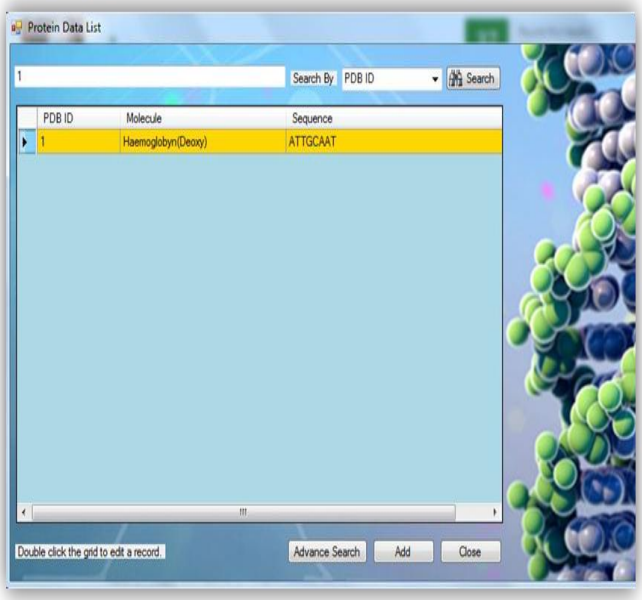


Fig. 5. Protein data retrieval using prototype model

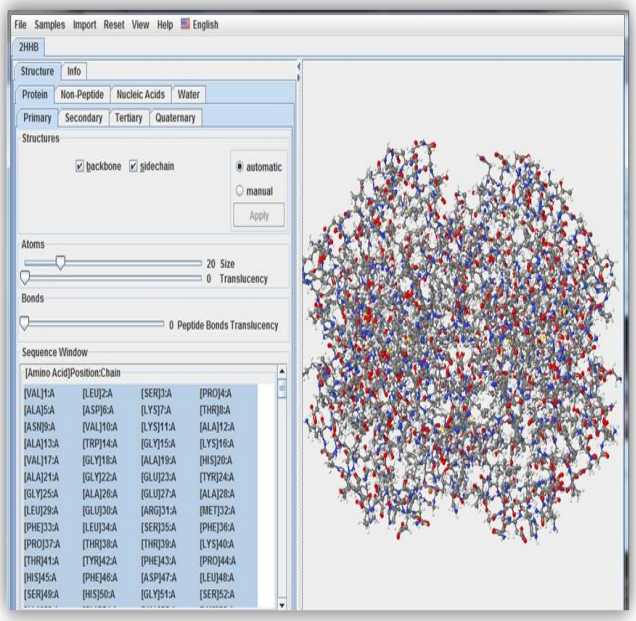


Fig. 6. Structural analysis of a protein

In the above fig. 5, a prototype model is shown, portraying how the software might actually work and search offline for already viewed data in the BLAST webpage. In fig. 6, the software is used to display the structural details and the information regarding it which is cross-referenced with the Protein Data Bank in order to provide the entire information regarding the protein.

IV. EXPERIMENTAL ANALYSIS

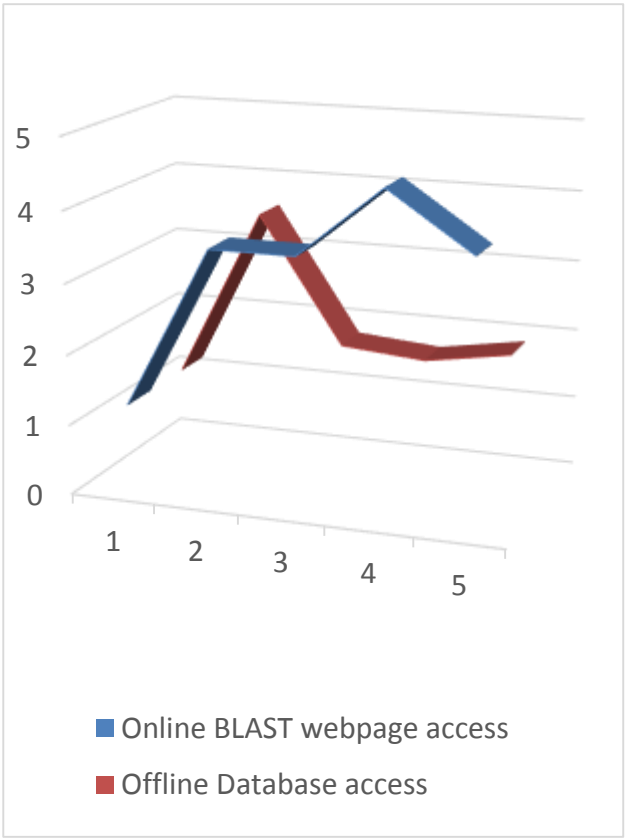


Fig. 7. Graphical Analysis showing difference in access time of data

According to the above graph stated in fig-7, a clear distinction in the difference in access time is shown. As per analytics, while the protein sequence is searched on the BLAST webpage, although being optimized, takes up time as the algorithm goes through the entire database for fetching the requested information. But if the accessed data is stored in the offline database, one can view its details much faster and is provided with all the information required as the data is cross referenced with Protein Data Bank.

V. CONCLUSION

This algorithm provides a new dimension to the workflow in the field of DNA sequencing. The algorithm provides an efficient caching of the data regularly viewed by the user in the BLAST web page into a personalized database providing a faster data access along with a high tone of security by encrypting the data before storing it into the database. Our proposed workflow is currently based on the BLAST tool and Protein Data Bank for data referencing, but it can be implemented with other sequencing tools as well and the encryption technique can also be replaced by other techniques as per the demand of the situation. The personalization of the study and the analytics of biological data is just a foundation and can be further modified by adding beneficial functionalities to the piece of software.

REFERENCES

[1] Jiang, J., 1996.Pipeline algorithms of RSA data encryption and data compression .In: proceeding IEEE International Conference on Communication Technology, 2:1008-1091

[2] Lian, S., J.Sun,Z. Wang and Y. Dai ,2004.A fast video encryption technique based on chaos .In: Proceeding the 8th IEEE International Conference on Control , Automation ,Robotics and Vision, 1 :126-131.

[3] Aikawa, M., K. Takaragi,S. Furuya and M. Sasamoto, 1998.A light weight encryption method suitable for copyright protection .IEEE Trans. consumer Electron, 44:902-910.

[4] Eagle, N. and A. Pentland, 2005. Social Serendipity: Mobilizing social software .IEEE Pervasive computing, pp: 4.

[5] Luke Alphey, 1997. DNA Sequencing: From Experimental Methods to Bioinformatics.

[6] GilAlterovitz and Marco R. Ramoni. Systems Bioinformatics: An Engineering Case-Based Approach.

[7] Raphael, Ben, Tang, Jijun, 2012.Algorithms in Bioinformatics**.** In 12th International Workshop, WABI2012, Ljubljana, Slovenia. Lecture Notes in Computer Science, Vol. 7534, (Eds.).

[8] Darbeshwar Roy, Bioinformatics.

[9] Nicholas H. Bergman. Comparative Genomics.