A web based nucleotide sequencing tool using BLAST algorithm

Ipsita Saha[#1], Joy Dewanjee[*2]

[#]*Assistant Professor, Computer Science and Engineering Department, Guru Nanak Institute of Technology*
Address Including Country Name

Abstract — *One of the most important methods which specifies the biological functionalities of a gene is by running a sequencing or similarity search on the existing protein and DNA sequence databases. The most famous tool for performing nucleotide sequencing is the NCBI's Basic Local Alignment Search Tool (BLAST) or its different various. This web based tool provides an alternative to the present BLAST tool, providing a faster access to the analytics of the protein structure with a more interactive interface. The tool also provides a user friendly and hassle free access with sorted and visually informative analytics.*

Keywords — *DNA, Protein sequence, Global sequence alignment, Homology, Basic Local Alignment Search Tool (BLAST).*

I. Introduction

With the increasing trend of research in the field of biotechnology and genomics, the researchers are facing many problems. Similarity has both a quantitative and qualitative aspect, where the similarity defines the quantity by the degree of similarity in the sequences and the alignment defines the quality by the mutual arrangement of the sequences being compared. Thus these aspects are highly impacted with increasing complexity of the biological data and increasing inherent information present in the sequences. This problem is mainly noted while representation of this biological data in terms of computational logic. The challenges faced while this implementation is widely based on the management of the biological information along with the algorithm for processing the data. Nowadays, in terms of performing a similarity search, the most popular tool is basic local alignment search tool (BLAST). The BLAST algorithm is a heuristic approach which indicates that it relies on some shortcuts to perform the search faster. BLAST performs "local" alignments. Most protein structures are modular in nature, with functional domains which are being repeated within the same protein as well as across different proteins from different species. The BLAST algorithm is implemented in such a way so that it could find these domains or shorter stretches of sequence similarity. If instead BLAST started out by attempting to align two sequences over their entire lengths (known as a global alignment), fewer similarities would be detected. The entire workflow is described in Fig. 1.

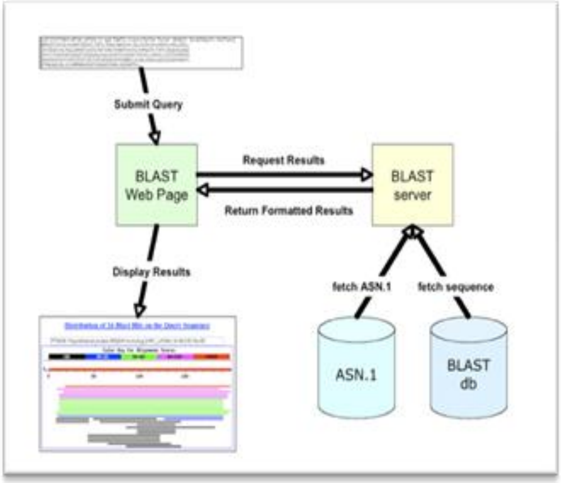


Fig. 1 Workflow of BLAST tool

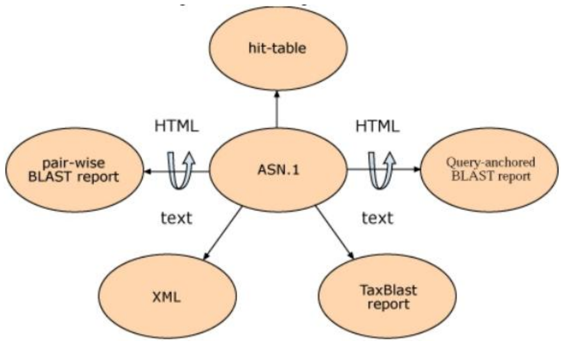


Fig. 2 Output formats from BLAST tool

There are drawbacks to parsing both the BLAST report and even the simpler hit table. There is no way to automatically check for truncated or otherwise corrupted output in cases when a large number of sequences are being screened. (This may happen if the disk is full, for example.) Also, there is no rigorous check for syntax changes in the output, such as the addition of new features, which can lead to erroneous parsing. Structured output allows for automatic and rigorous checks for syntax errors and changes. Both XML and ASN.1 are examples of structured output in which there are built-in checks for correct and complete syntax and structure. (In the case of XML, for example, this is ensured by the necessity for matching tags and the DTD.) For text reports, there is often no specification, but perhaps a (incomplete) description of the file is written afterward. The output format scenarios are depicted in Fig. 2.

A. *Requirement of sequencing tool*

In order to implement a generic DNA sequencing tool, the basic ideology which is involved is that it would involve a huge number of DNA sequences. The implementation of homology searches of the sequences is generally the initial step of any sequencing tool. But while preparation of the logic of implementation, most people ignore the importance of the computational logic and the design of an optimised algorithm in order to handle such bulk amount of data.

B. *Aim of implementation*

Thus keeping the basic ideology of implementation in mind, we have formulated our sequencing tool which is capable of handling the huge amount of biological data and at the same is able to manage the data efficiently so that the retrieval time from the database is minimal with accuracy. Our tool also focuses on a user friendly and visually informative interface to ease the process of searching for the user.

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. Apart from the above mentioned work, other web based tools which were taken into consideration while implementation of our tool were WebBLAST 2.0 [3] and OCGC BLAST [2]. Although both the tools offer an accurate result set with proper sequencing and analytical information, but they are based on the flat file based data file and does not incorporate any database functionalities.

III. Proposed Work

Our sequencing tool simplifies the complex analytical procedure behind the nucleotide sequencing by providing a simple and informative result set. The tool accepts input in the form of web based input or XML file based input and executes a detailed search referring to the underlying database and then provides a result set with the similarity details.

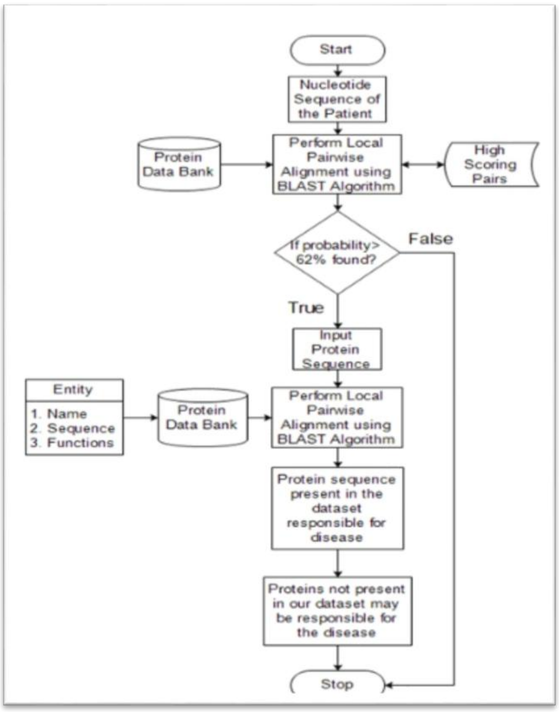

Start
Nucleotide Sequence of the Patient
Protein Data Bank
Perform Local Pairwise Alignment using BLAST Algorithm
High Scoring Pairs
If probability> 62% found?
False
True
Input Protein Sequence
Entity
1. Name
2. Sequence
3. Functions
Protein Data Bank
Perform Local Pairwise Alignment using BLAST Algorithm
Protein sequence present in the dataset responsible for disease
Proteins not present in our dataset may be responsible for the disease
Stop

Fig. 3 Workflow of sequencing tool

As describes in the Fig. 3, the workflow formulates a thorough homology search based on the database which is linked to the standard biological database such as NCBI databases, etc. Due to the implementation of buffer logic, the execution time of the algorithm increases with each iteration no matter what be the sequence length.

A. *Architectural Overview*

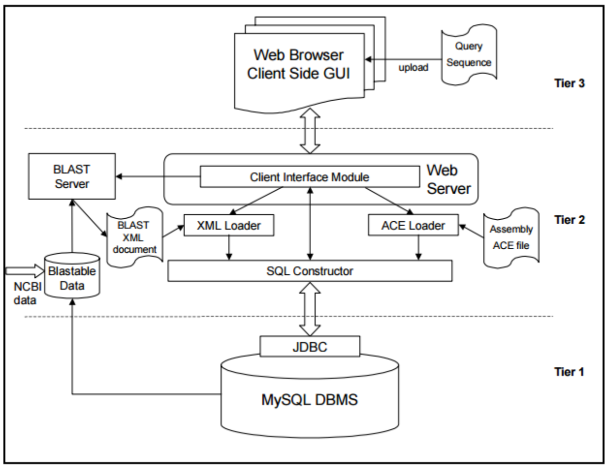


Fig. 4 Basic architecture of proposed sequencing tool

As the Fig. 4 depicts our tools is designed, based on a 3-tier architectural model where each tier corresponds to a functionality of the tool.

In the third tier or the Presentation Layer, it contains the implementation of the client interface. The user interface is constructed in dynamic and responsive web pages which is viewable through any form of devices as the pages are capable of adjusting depending on the device screen width. He layout was developed in PHP environment. The Client-side processing basically includes the responsive representation of the output in HTML format, validation for user input after which the request is passed further for processing.

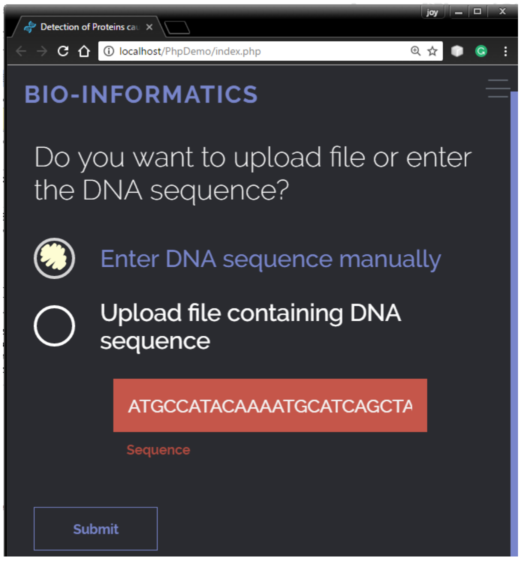
The second tier or the conceptual layers contains the brain of the tool. It consist of the multi-threaded JAVA application which is responsible for major operations like:

- The client portal acts as an interfacing unit between the Presentation layer and the conceptual layer for exchange of data.
- The Loader functionality for uploading the data from the presentation layer in XML format and then parses it for further processing.
- Matching of the sequence with BLAST searches against NCBI as well as the internal database of the application.
- The buffer module for decreasing the access time of the retrieving logic for the sequences.
- Processing of the input sequence for proper alignment using local alignment algorithm.
- An SQL constructor is implemented which acts as an interfacing unit between the conceptual and the Database layer.

The first tier or the database layer consists of the actual database of the application for storing the biological information regarding the protein structures and their relative information. The database also stores the sequence search results and their related information such as hit definition, expect value, bit score, gap scores pairwise alignments and so forth. For each query sequence submitted, the information is referenced with internal database along with the international standard databases.

B. *User Interface*

The sequencing tool breaks down the complex process of sequence analysis with the means of easily accessible information and detailed data visualisation. The tool has the capability of maintaining separate database for each of the user using the tool and keep a tab of their research work. Apart from this, the main focus of the tool is similarity search, which is triggered by either accepting a DNA sequence manually or the user can upload a file in case it's a large file as depicted in Fig. 5. After the submission of the input sequence, the PHP pages does the initial validations of the sequence and passes it to the actual brain of the tool. By actual brain, we refer to the multi-threaded JAVA application which receives the data from the PHP pages and does the similarity search and sequencing. While the sequencing process, the NCBI database is referred at the runtime if the sequence is new to the system. After this the similarity sequencing result sets are stored in the internal database for future reference. A copy of the result set is also kept in the buffer memory in order to reduce the search cost from the next iterations. After all the processing is complete, the user is able to see the result in the PHP pages with similarity percentages along with the detailed sequencing information and biological data related to it as depicted in Fig. 6.



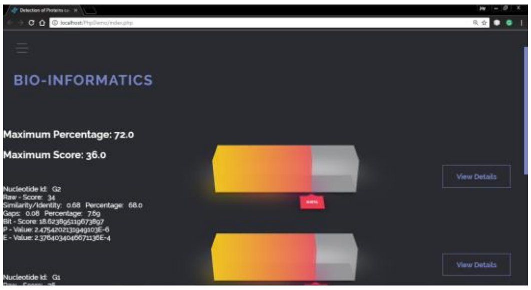
Do you want to upload file or enter the DNA sequence?

● Enter DNA sequence manually

○ Upload file containing DNA sequence

Sequence: ATGCCATACAAAATGCATCAGCTA

Fig. 5 Input interface of sequencing tool



Maximum Percentage: 72.0

Maximum Score: 36.0

Nucleotide Id: G2
Raw - Score: 34
Similarity/Identity: 0.68 Percentage: 68.0
Gaps: 0.08 Percentage: 7.69
Bit - Score: 18.623895119673897
P - Value: 2.4754202131949103E-6
E - Value: 2.3764034046671136E-4

Nucleotide Id: G1

Fig. 6 Result set details of sequencing tool

IV. EXPERIMENTAL ANALYSIS

We did an extensive test on different data sets which would correspond to the performance of the tool and the result was quite promising. After a few database hits, the data access time decreased a lot due to the buffer logic implementation.

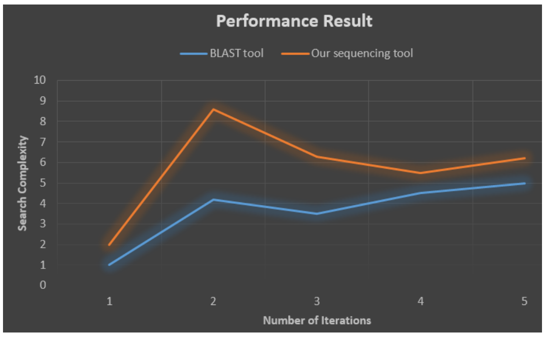


Fig. 7 Experimental Graph

As compared to the execution of the BLAST tool, the execution time of our tool was considerably faster. As provided in the Fig. 5, it is seen if the sequence provided as input is small in length both our tool works almost same as the BLAST tool. But as the length of the sequence increases thus increasing the search complexity and the number of calls are repeated, it is seen that the execution time of our tool decreases and the retrieval rate is faster. Although the initial execution time of our tool is high, but later due to buffer logic, the performance increases.

V. Conclusions

We have described our web based sequencing tool as an interactive and visually informative alternative for the present BLAST tool based on the local pair wise alignment. We were successful in importing and storing structural data of protein and nucleotides keeping the paradigms of relational database as guide. The retrieval and analytical description of the protein sequences was quite optimized and faster due to implementation cache storage technique. As a future scope to our tool, we are planning to enhance the tool by implementation of global alignment sequencing technique which will result in an integrated tool for sequence alignment with optimized result.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," Journal of Molecular Biology, vol. 215, pp. 403-410, 1990.

[2] J. Cuticchia, S. Parameswaran, R. Alexandrova, and E. Crowdy, "OCGC BLAST," http://www.ocgc.ca/ocgcblast.htm., 1999.

[3] E. S. Ferlanti, J. F. Ryan, I. Makalowska, and A. D. Baxevanis, "WebBLAST 2.0: an integrated solution for organizing and analyzing sequence data," Bioinformatics, vol. 15, pp. 422-423, 1999.

[4] Glass JI, Hutchison CA 3rd, Smith HO, Venter JC, “A systems biology tour de force for a near-minimal bacterium.”, Mol Syst Biol, pp. 5-330, 2009.

[5] Clatworthy AE, Pierson E, Hung DT, “Targeting virulence: a new paradigm for antimicrobial therapy”, Nat Chem Biol, pp. 3:541–548, 2007.

[6] Furney SJ, Alba MM, Lopez-Bigas N, “Differences in the evolutionary history of disease genes affected by dominant or recessive mutations.”, pp. 7-165, BMC Genomics 2006.

[7] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, LucauDanila A, Anderson K, Andre B, “Functional profiling of the Saccharomyces cerevisiae genome Nature”, pp. 418,387–391, 2002.

[8] Cullen LM, Arndt GM, “Genome-wide screening for gene function using RNAi in mammalian cells”, Immunol Cell Biol, pp. 83,217–223, 2005.

[9] Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, “Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiology”, pp. 50,167–181, 2003.

[10] Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW, “Evolutionary rate in the protein interaction network.”, Science,pp. 296,750–752, 2002.

[11] Jordan IK, Rogozin IB, Wolf YI, Koonin EV, “Essential genes are more evolutionarily conserved than are nonessential genes in bacteria”, Genome Res, pp. 12,962–968, 2002.

[12] Green ED, Guyer MS., “Charting a course for genomic medicine from base pairs to bedside”, Nature 470 (7333):204-213. doi: 10.1038/nature09764, 2011.

[13] Needleman SB, Wunsch CD., “A general method applicable to the search for similarities in the amino acid sequence of two proteins”, J Mol Biol 48 (3), pp. 443-453, 1970.

[14] S.C. Rastogi, Namita Mendiratta, “Bioinformatics concepts, skills and applications”, pp. 124-250, 2006.

[15] Bryan Bergeeron, “Bioinformatics computing”, pp. 330-335, 2010.

[16] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/.