International Journal of Computational and Experimental Science and ENgineering (IJCESEN)
Vol. 5-No.2 (2019) pp. 76-79
http://dergipark.org.tr/ijcesen

ISSN: 2149-9144

Research Article

A Framework for Query Optimization Algorithms for Biological Data

Khalid Mohammad JABER*, Nesreen A. HAMAD and Fatima M. QUIAM

Al-Zaytoonah University of Jordan, Department of Computer Science, 130 Amman 11733 Jordan.

* **Corresponding Author** : k.jaber@zuj.edu.jo
ORCID: 0000-0002-8458-401X

Article Info:

DOI: 10.22399/ijcesen.508889
Received : 06 January 2019
Accepted : 10 June 2019

Keywords

Query Optimization
Searching Algorithms
DNA
Protein
Bioinformatics
Data Mining

Abstract:

Recently, the size of biological databases has significantly increased, with a growing number of users and rates of queries. As a result, some databases have reached a terabyte size. On the other hand, the need to access the databases at the fastest possible rates is increasing. At this point, the computer scientists could assist to organize the data and query in a way that allows biologists to quickly search existing information. In this paper, a query optimizing model for DNA and protein sequence datasets is proposed. This method of dealing with the query can effectively and rapidly retrieve all similar proteins/DNA from a large database to reduce the computational time required to search these biological genomic datasets. Specifically, a theoretical and conceptual proposed framework is derived using query techniques from different applications. The results show that the query optimization algorithms reduce the query processing time in comparison with the normal query searching algorithm.

1. Introduction

Biological databases are stores of biological information that are collected from various scientific experiments and published literature. Meaningful knowledge and information can be extracted from these databases such as: similarities of biological sequences and structures, gene function, structure, and other useful information [7]. Examples of nucleotide and biological sequence databases that can be downloaded for free are: the GenBank databases [1], the genomic laboratory database of European Molecular Biology [2], [3], and the Databank of DNA in Japan [4].

One of the first databases to arise was GenBank, which is a collection of all available protein and DNA sequences. The DNA primary structure is represented by a single linear sequence where the DNAprimary-sequences consist of four nucleotide bases, which are represented by the following alphabetical characters: DNA primary-sequence *{ A, C, G, and T }*. Due to the large length of the real DNA single sequence; the FASTA format is used to store these sequences in the genomic databases to be more readable [5], [6]. Figure 1 illustrates an example of the DNA primary structure, as part of Homo sapiens chromosome1 from GenBank database. While, "N" alphabet character is used as neutral DNA nucleotide base to represent the unknown part and so called wild-card characters. The second genomic data in GenBank is the protein, which are polypeptide chains [8]. In these databases the protein could be a combination of 20 amino acids, which are listed in Table 1. Having more characters result in making the structure of protein more complex than the DNA primary structure. The rapidly increasing amount of data, number of users and rates of queries is resulting in databases with huge sizes, which is a challenge


>gi|157697810|ref|NW_001838472.1| Homo sapiens chromosome 18 genomic contig, alternate assembly, whole genome shotgun sequence

GTGGTGATGGTGATGGTGATGATGGTGGTGGTGATGGTGATGGTGATGATGGTGGTGATGGTGGTAGTGAT
GGTGGTAATAGTGATTCTGAACTGCCTCTTAGGGGCTTAGGGGAAAGCAATGTGCAGAGTTGAAGTCACCTC
AGCAGGGTGCCTGGAGATGAGACCAGTAGAAATACAGAAGTGATAAACAGAGCCTCAGGGGAAGGTTCTCC
ACGTAAGTAGAATGACCAATCTCCAAAAGAATTCATTGGCATTGAAACAGATTCTAACTGTTCAGAATTTTCTA
AACGTTTCTCAAAAATGAAAATTCTCCCACTTGGGAAGGCAGCAGAGAAGAGGAGAAACAGC

Figure 1. *DNA Nucleotide Primary Structure of Homo Sapiens Chromosome1 in FASTA format.*

Table 1. *Amino ACIDS Abbreviations [14]*

Amino Acid	3 Letters	1 Letters
Alanine	Ala	A
Arginine	Arg	R
Aspartric acids	Asp	D
Asparginine	Asn	N
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycime	Gly	G
Histine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methione	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

for the computer science researchers. In particular, for any given query in a large database, the need to quickly retrieve all data which are similar is essential but considered as a difficult and time-consuming task. In this paper, the focus will be in the increasing amount of data in biological databases. An example of this large increase can be found in the DDBJ, EMBL and GenBank from 1999 to 2016, its exponentially growth is presented in Figure 2, which is adapted from [9]. Such huge databases cause the problem of the computational search that is required to extract data. Therefore, this paper proposes to assist in searching existing biological databases in an easy, quick and effective way by focusing on the query optimization algorithm to reduce the computation time for biological genomic datasets searching algorithm.

This paper is organized as follows. Section 2 covers some of the related works. The main method is presented in Section 3. The benchmark is discussed in Section 4. Finally, Section 5 provides the conclusion remarks along with a list of the future works.

2. Related Works

In this section, some of the related works that applied query optimization algorithm to tackle the problem of searching and extracting data are presented. Korf et al. [10] defined the Query multiplexing or packing as a mechanism

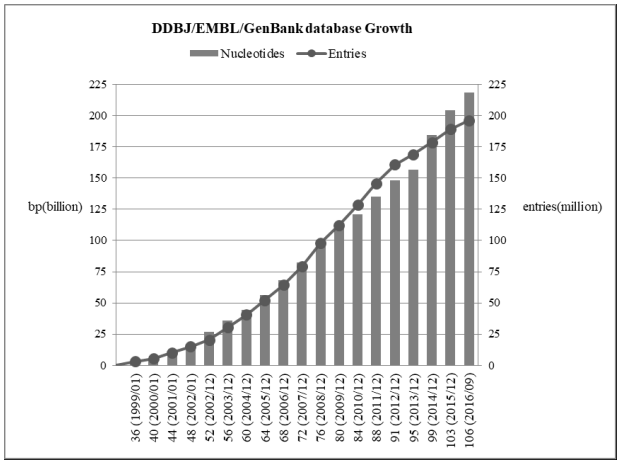

Figure 2. *The exponential growth of the DDBJ, EMBL and GenBank from 1999 to 2016 (Adapted from [15]).*

to pack multiple queries together and reduce the overhead of reading the queries repeatedly. An example was given to explain this idea: if there is a collection of 100,000 sequences from a favorite organism, where the goal is to search this against all other organisms in the public database. If 100,000 searches are executed and the search is one at a time, then 100,000 times is needed to read this sequence. However, by using query multiplexing the searching performance can increase [10]. Moreover, there is a version from BLAST searching algorithm that deals with query packing which is called MegaBLAST. MegaBLAST is a specialized version of BLASTN that supports multiplexes queries [10]. Additionally, query chopping is another mechanism to increase the performance and is used for larger query sequences where more memory is required when searching. In query chopping, the individual query sequence is divided into several segments, searching them separately and then the results are merged back together. However, dealing with alignments that cross the boundaries between segments is the main difficulty with query chopping [10]. To address this limitation (Rieffel et al. [11]) proposed Paracel BLAST, that takes a different codebased approach that directly handles the boundaries between segments and reprocesses them as appropriate.

Query optimization is not only used in biological databases, but also can be used in online analytical processing (OLAP) such as Starburst and Volcano. The IBM DB2 optimizer depends on Startburst, while the Microsoft SQL-Server optimizer is based on Volcano. The main difference between the two approaches is the method in which alternative plans are generated. Starburst generates the plans bottom-up. On the other hand, Volcano generates the plans top-down. More details can be found in [12]. In a previous work, Jaber et al. [16] proposed a framework of the decision tree indexing model

(DTIM) to build indexes for enormous DNA-protein datasets to reduce the computation time for searching algorithm, as well as reducing the space required to do the computation and store indexes.

3. Query Optimization for Biological Data: Framework

Query optimization is deployed in this study to solve the problem of searching and extracting meaningful information from the biological data in GenBank. Due to the limitations of the queries reading approaches and also the increased length of query sequences which require more memory to search, query optimization is adapted in this paper to enhance the query pattern in the whole datasets to solve the problem of extracting biological data.
To clarify the problem area, Figure 3 shows the main steps of the proposed framework in general. In the subsequent subsections, the phases are stated in detail. The main steps include:

1) Scanning, parsing and validating phase.
2) Query optimizer.
3) Searching queries in datasets.

3.1. Scanning, Parsing and Validating

The aim of this phase is to prepare the representation of the queries. This phase is carried out in three steps. The first step is to scan the queries tokens from the text file by opening the file, and reading the containing data line by line. Each line is filtered to remove the wild-card characters such as N and X. The second step is to parse and build the query representations which are called Text (**Q**). The queries are represented as a collection of rows/sequences (*S*) that may or may not contain duplicates. Each row is represented as a set of character values (V) = $\{v_1, \ldots, v_i; \ldots, v_n\}$, where ($v_i$) = {*A, T, G, and C*} shows all the possible values of data Q. In the third step, these built queries representations are validated and stored in the database as a multiple flat file dynamically and based on the first character of each row.

3.2. Query Optimizer

The query multiplexing or packing mechanism is used in this phase to handle multiple queries at the same time. To handle query multiplexing, the control flow statement that enables any number of instructions to be executed repeatedly based on a given condition is used. Therefore, by using the control flow statement, the searching algorithm is repeatedly executed based on the number of queries. Additionally, to enhance the searching

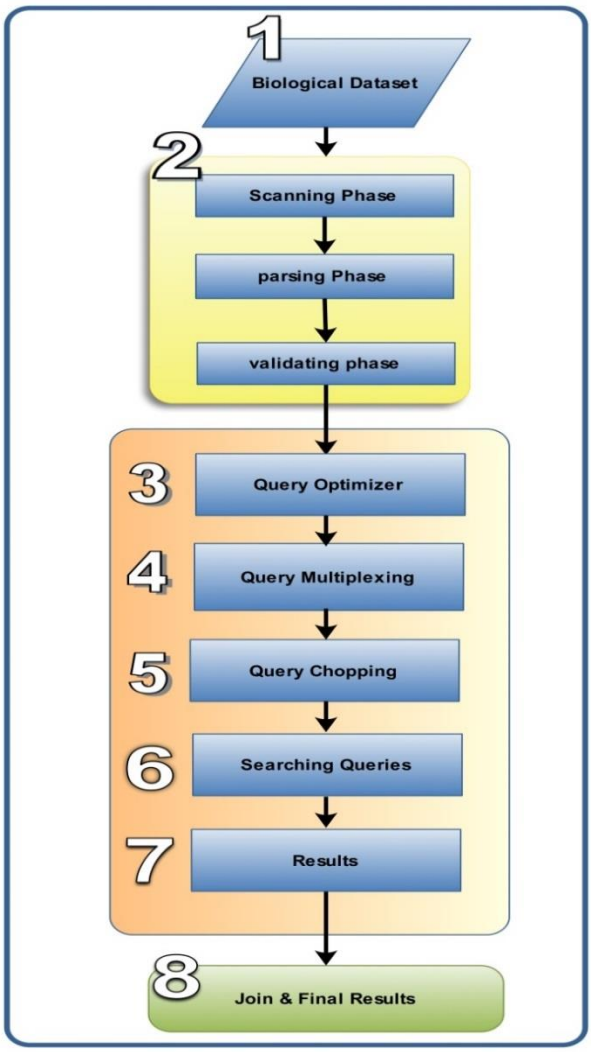

Figure 3. *Workflow for Query Optimization*

speed, the query chopping technique was adapted to merge in this phase. After adapting the query chopping for the proposed method, the individual query sequence are divided into several segments, use particular segments to search them dependently and then show the result. Figure 4 presents the query sequence, which is divided into five segments and the segments needed for the query process are only a_5, a_1 and a_2, whereas segments a_3 and a_4 are not necessary in the query search process. For example, suppose that there is a query of 80 nucleotides length, only 50 nucleotides in the search process are used. Therefore, instead of searching all the 80 Nucleotides, the use of a particular nucleotide increases the speed of the search.

3.3. Searching Queries in Data sets

The third phase of this framework involves the searching process. The searching process of this work can be used to search the queries in the indexing DNA Database. Afterwards, the searching results are all joined to get the final results.

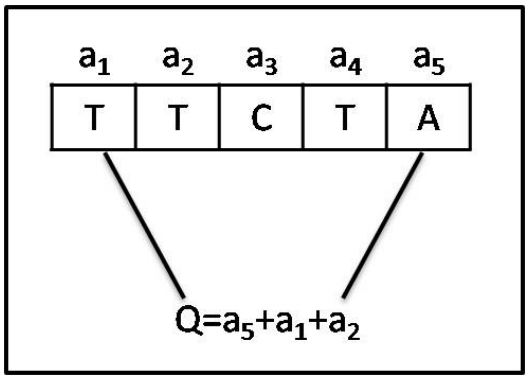

Figure 4. *Query Chopping Technique*

4. Benchmark

To evaluate the proposed method, different tests will be performed and compared with BLAST+ which can be downloaded from (NCBI Website, [13]) in order to test the query processing time(elapsed time), which is the time spent on finding all the queries, which are submitted to the proposed model or BLAST+. The effects of variant data sets sizes will also be examined, with changing the number of queries.

5. Conclusion

This paper introduced a query optimization method to enhance the query processing time in huge chunk of biological data in GenBank. The proposed query optimization method is based on the query multiplexing or packing mechanism and query chopping algorithm. This method can effectively and rapidly retrieve all similar proteins/DNA from a large database. A theoretical and conceptual proposed framework is derived using query techniques. This method can be developed further by applying Artificial Intelligence (AI) computational methods such as Particle Swarm Optimization (PSO), and Harmony Search (HS). These AI tools can be incorporated in the query optimization method to support its implementation in the future.

Acknowledgement

The authors would like to sincerely appreciate the useful and insightful comments from the respectful and anonymous referees. These comments have been greatly used to improve the transcript of the paper and to clarify the presentation of the study. This research was funded by Al-Zaytoonah University of Jordan.

References

[1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2010). Genbank, Nucleic acids research. 38(1), D46-D51, DOI: 10.1093/nar/gkx1094.

[2] P. Rice, I. Longden, and A. Bleasby (2000). Emboss: the european molecular biology open software suite. Trends in genetics, 16 (6), 276-277. DOI: 10.1016/S0168-9525(00)02024-2

[3] A. Bairoch and R. Apweiler (2000). The swiss-prot protein sequence database and its supplement trembl in 2000. Nucleic acids research, 28 (1), 45-48. DOI:10.1093/nar/28.1.45

[4] K. D. Pruitt, T. Tatusova, and D. R. Maglott (2007), Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic acids research. 35(1), D61-D65. doi: [10.1093/nar/gkl842]

[5] P. Librado and J. Rozas, Dnasp (2009). v5: a software for comprehensive analysis of dna polymorphism data, Bioinformatics. 25(11), 1451-1452. DOI: 10.1093/bioinformatics/btp187

[6] C. Plot (2000). The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences, Biotechniques. 28(6), 1102-1104. DOI:10.2144/00286ir01

[7] Jaber, K. M., Abdullah, R., and Rashid, N (2014). A. Fast decision tree-based method to index large DNA-protein sequence databases using hybrid distributed-shared memory programming model. International Journal of Bioinformatics Research and Applications. 10(3), 321-340. doi: 10.1504/IJBRA.2014.060765.

[8] R. J. Block, D. Bolling et al. (1945). The amino acid composition of proteins and foods. analytical methods and results. The amino acid composition of proteins and foods. Analytical methods and results. 17(4).

[9] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson et al. (2011). The european nucleotide archive, Nucleic acids research. 39, D28-D31.

[10] Ian Korf, M.Y., Joseph Bedell (2003). BLAST.

[11] Rieffel, M. A., Gill, T. G. and White, W. R. (2004). Bioinformatics clusters in action., Cluster World.

[12] Prasan Roy(2000). Rule-Based Query Optimization using the Volcano Framework., PhD thesis, IIT Bombay.

[13] NCBI Website, URL: http://blast.ncbi.nlm.nih.gov,2018.

[14] Whitford, D., Proteins (2005). Structure and Function., 1 Edition, Wiley, 2005.

[15] DDBJ Database Available at: http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-old-e.html. [Accessed 12 April 2017].

[16] Khalid Mohammad Jaber, Rosni Abdullah and Nur'Aini Abdul Rashid. Fast Decision Tree-Based Method to Index Large DNA-Protein Sequence Databases Using Hybrid Distributed-Shared Memory Programming Model. International Journal of Bioinformatics Research and Applications, Volume 10, No. III, pp. 321-340, 1 January, 2014.