International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Similarity Analysis of DNA Sequences Using BWT Technique

Syed Asma Andrabi[1], Dr. Manoj Kumar Gupta[2]

[1]M.Tech CSE (S/W) Associate Professor
[2]SET, Sharda University SET, Sharda University

Abstract: ***DNA sequences similarity comparison is important for analyzing genetic basis of organisms.There is a repetition of some patterns over a length of sequence. It contributes to high computational cost, maximum memory usage and less performance while comparison. A requirement to optimize DNA sequences matching with quick access and accuracy is needed. A number of techniques have been developed so far to compare patterns of DNA sequences with reference sequences. These compare sequences with high retrieval time, hence less efficiency while comparison and location can't be detected. To access DNA bases in a DNA database and increase efficiency with fast search and exact location BWT technique is implemented on DNA database.***

Keywords: BWT, DNA Sequences, Query Comparison, Genetic Testing

1. Introduction

DNA sequence analysis has various applications in biological domain like Molecular biology which is branch that studies the genome itself, how proteins are made, what proteins are made of, identifying new genes and associations with diseases and phenotypes, and identifying potential drug targets. Other field is of Evolutionary biology which studies how different organisms are related and how they evolved.

Met genomics is other area where DNA Sequence analysis has its application i.e. identifying species present in a body of water, sewage, dirt, debris filtered from the air, or swab samples of organisms. It is also helpful in ecology, epidemiology, microbial research and other fields. Less-precise information is produced by non-sequencing techniques like DNA fingerprinting. This information may be easier to obtain and is useful for detecting the presence of known genes for medical purposes (see genetic testing), Forensic identification, parental identification. Therefore there arises a need to develop such methods which can analyse the DNA sequences similarity in a quicker, efficient and accurate way.

2. Problem Statement

The DNA similarity analysis has been a matter of concern over the years for analyzing the genetic basis of the organisms. The algorithms have been developed for the same but a precision is what that had been missing in the previous algorithms. The algorithms developed in the past were not that much accurate and fast to analyze the DNA sequences.

3. Proposed Methodology

Similarity analysis of DNA sequences is important for inferring the genetic basis of organisms. So far many techniques and methods have been developed to compare patterns of DNA sequences with reference sequences. There is a repetition of some patterns over a length of sequence. It contributes to high computational cost, maximum memory usage and high retrieval time while comparison. There lies a requirement to optimize DNA sequences searching and matching with efficient search method. Therefore, this kind of problem needs to be addressed with a more efficient technique.

BWT technique is implemented on DNA database and sequences are compared to generate results quickly. An accurate Position of bases is found. All sequence analysis tasks can be accomplished with a suffix array which takes less space and less time. Suffix array indexes bases in a database.

Initially in pre-processing phase DNA (De oxy Ribo nucleic Acid) database is processed with a BWT (Burrow Wheeler Transform). Earlier some techniques were used to perform comparison without any processing on data and consumed more time.

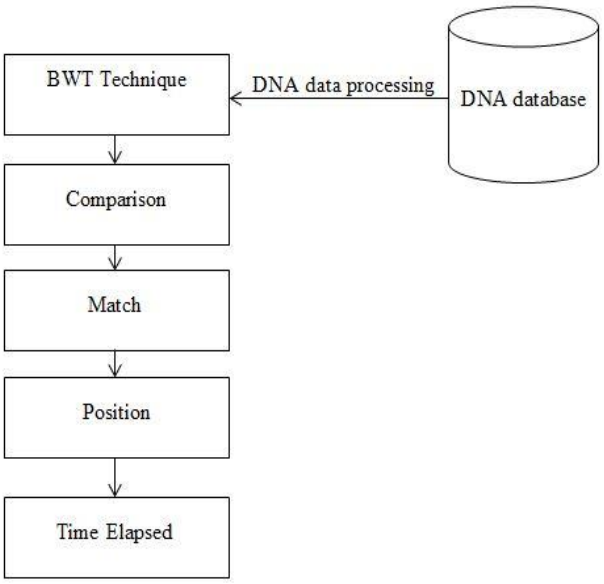

Figure 1: Proposed Technique Blocks

The DNA databases are tested through the BWT technique at first instance and then the DNA sequences are compared as mentioned in the above steps. The time elapsed for the similarity analysis of the proposed algorithm yields better results than the earlier ones(techniques).

Here a BWT (Burrow Wheeler Transform) is applied on DNA database. A DNA database is filtered by this technique for further processing. A sample DNA sequence is then compared with the processed database. After the comparison is drawn by checking out whether a sample sequence matches or not, further processing is done to find out the position of bases in a database. For finding out a position search algorithm is applied on a database on which BWT has been implemented and time elapsed for different samples against database is calculated to find out that result is retrieved quickly.

4. Results and Discussion

The final results of the proposed algorithm have been shown in the tabular form below:

Table1: Size: 4723, Bytes: 9446

Sample	Ggtgcgcacacgagaag(Q1)	ccagcgctcttg(Q2)	cgcacacgagaaggacgcgcggcccccagcg(Q3)	cgaaagaaaaaaa(Q4)
Length	17	12	31	13
Match	1	1	1	No match
Start	22	51	26	
End	38	62	56	

The table for the total time elapsed for the similarity of DNA sequences is mentioned in the table shown below:

Table 2: Time Elapsed

Time(s)	Q1	Q2	Q3	Q4
T1(BWT)	0.602634	0.49834	0.971702	0.357037
T2(Pairwise)	5.662951	5.336202	5.850669	5.042558

Table 3: Size: 14169 Bytes: 28338

Sample	ggtgcgcacacg(Q1)	aaaaaaattttg(Q2)	cgcacacgagaaggacgcgcggcccccagcg(Q3)	cgaaagaaaaaaaa(Q4)
Length	12	12	31	13
Match	3	No match	3	No match
Start	22,4745,946		26,4749,9472	
End	38,4756,9479		56,4479,9502	

Table 4: Time Elapsed for Proposed Database

Time(s)	Q1	Q2	Q3	Q4
T1(BWT)	1.318336	0.598934	3.168392	1.004642
T2(Pairwise)	6.662951	5.836202	7.745408	5.042558

The proposed algorithms make use of BWT technique which has reduced the time elapsed to generate the DNA similarity sequences. The BWT technique has a significant role in the reduction of time and it has been proved by comparing the results of base paper and the proposed algorithm. BWT is a technique which when implemented on DNA database assists in retrieving sample sequence query quickly from the same while comparing query sequence in DNA database. Fast retrieval of sample or query sequence occurs because BWT technique compresses the huge DNA database. The implementation has been done in MATLAB.

5. Analysis

Comparison of query with a DNA database is as follows:

Firstly a sample sequence query comparison is drawn with DNA database using BWT approach. Sample sequence queries e. g, Q1,Q2,Q3,Q4 are taken as shown in table 1 and table 3 and the time in which query is matched is calculated as shown in table 2 and table 4.

Secondly, a sample sequence query comparison is drawn with DNA database where pair wise approach has been applied. Sample sequence queries e.g., Q1, Q2, Q3, Q4 are taken as shown in table 1 and table 3 and the elapsed time in which query is compared is calculated as shown in table 2 and table 4.

These comparisons are done to analyse the time taken by two approaches. The BWT technique takes less time for retrieval of sample sequence queries.

An efficient approach to reduce DNA sequence comparison time from a DNA database with quick access and accurate location is depicted.

In this experiment, for different sizes of sample queries the implemented technique outperforms traditional approaches. The DNA database is stored in compressed form which requires less computational overhead to answer select queries. The BWT technique implementation performs significantly better than the traditional methods. An efficient approach to reduce DNA sequence comparison time from a DNA database with quick access and accurate location is depicted.

6. Discussion

Results have been shown by comparing sample sequence with BWT implemented DNA database and time interval is calculated in which query is compared. In next step, sample sequence is compared with DNA database and time is calculated. After the two techniques have been implemented, time intervals are calculated for the respective techniques, and position of bases is determined.

Results have been summarised in the table given above:

Sample sequence queries are tested for comparison and their length, starting position, end position is determined, if a match is found then only it could proceed for further processing like searching the DNA database, as shown in table1 and table3.Time required for comparison of a sample sequence is calculated for techniques implemented on DNA database as shown in table 2 and table 4.Two techniques shown are BWT implemented on DNA database and pairwise alignment.

Compare the query sequence against a large amount of positions in the DNA database. This is possible because the BWT technique has been implemented on DNA database. This has much computing power that may be utilized. The main idea was to use as simple as possible approach to compare sequences such that time consumed is less and positions are found in the DNA database most likely to give a fast retrieval with increased efficiency. And then apply the search occurrence algorithm to find the best among positions.

The goal of this project was to try to speed up retrieval using the BWT technique which compresses the data.

- The algorithm used does index lookups to find positions in the DNA database.
- The next iteration counts the bases and applies base count function and performs the index lookups number of bases to be checked.
- Sample query match is found using LFC function and calculates matched string distance length.
- Initialization is made that sample DNA sequence query can be located anywhere in reference DNA database.

A number of issues occur while dealing with a huge DNA database and major problem arises while comparison of sample sequence is to be made.AS DNA database comprises of data i.e. bases which are the basic expression of identity in billions. In other studies on similarity analysis of DNA sequences, many techniques have been used for the purpose of comparison. A graphical depiction is made for comparisons for two techniques as shown in figure 2 and figure 3.

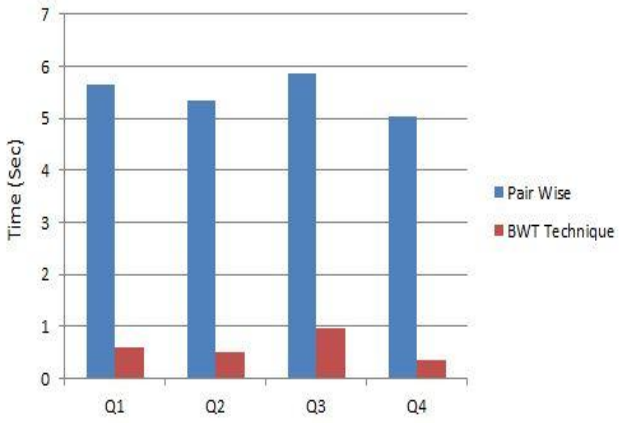

Figure 2: Comparison using BWT and pairwise approach

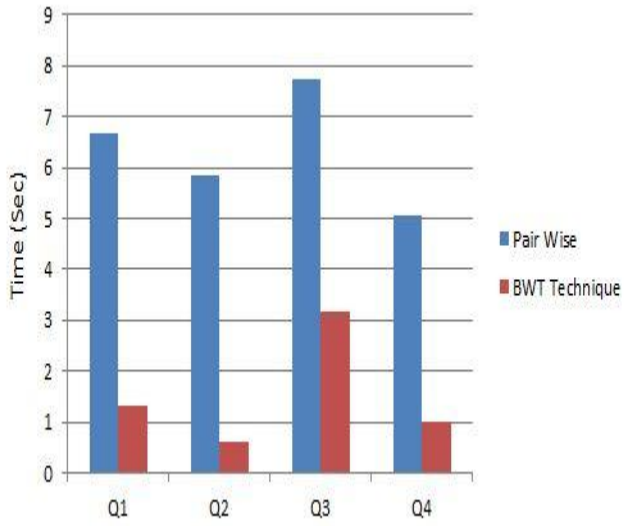

Figure 3: Comparison using BWT and pairwise approach

7. Conclusion

Here a reference genome is converted to an indexing data structure based on the Burrow Wheeler Transform (BWT), from which matches to individual query sequences can be rapidly determined with exact location in a DNA database. Therefore there lies a need for an approach to deal with growth of sequences in a DNA database which focuses mainly on storage, compression of the data. Overall performance comparisons can be viewed in many ways. The performance metric that was developed in this thesis shows a valuable indication of the processing ability while still maintaining an aspect of the architecture's power. Without this significant change, attempting to view the results from comparisons between 2 different architectures does not give the view needed to understand their

relationship. The result of a supercomputer's processing ability when compared to a laptop computer would show dramatically better performance when considering only the processing power. When the power consumed by the supercomputer is taken into account, the results are levelled out to understand that the laptop's performance may not be insignificant.

8. Future Work

BWT technique can further be extended for future use in coping up with the mutations i. e, deletions or insertions created in a DNA sample sequences .And the performance capacity is increased.

References

[1] Anthony J. Cox, Tobias Jakobi, Giovanna Rosone and Ole B,Schulz-Trieglaff, Comparing DNA sequence collections by direct comparison of compressed text indexes, Volume 7534, 2012, pp 214-224
[2] A. J. Cox, M. J. Bauer, T. Jakobi, and G. Rosone: Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform. Bioinformatics,Vol.no.2 2012,Pages 1–6
[3] Seon Jeong, Kyoung-Wook Park, Seung-HoKang, Hyeong-Seok Lim, An efficient similarity search based on indexing in large DNA databases, Computational Biology and Chemistry / Computers & Chemistry - CANDC , vol. 34, no. 2,2010, pp. 131-136
[4] Xia Cao Shuai Cheng Li Beng Chin Ooi Anthony K.H. Tung, Piers: An Efficient Model for Similarity Search in DNA Sequence Databases, Department of Computer Science, National University of Singapore, Vol. 33 Issue 2,2004, Pages 39-44
[5] Manoj Kumar Gupta, Rajdeep Niyogi, Manoj Misra, A New Adjacent Pair 2d Graphical Representation Of DNA Sequences, Journal of Biological Systems, Vol. 21, No. 1, 2013
[6] M. J. Bauer, A. J. Cox, and G. Rosone, Lightweight algorithms for constructing and inverting the BWT of string collections, Theoretical Computer Science, 2012.
[7] X. Chen ,A compression algorithm for DNA sequences and its applications in Genome comparison,Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, 2002
[8] P.Ko,S.Aluru ,Space efficient linear time construction of suffix arrays, Lecture notes in computer science 2676, 2003,200-210
[9] D. Adjeroh, T. Bell, A. Mukherjee, The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. Springer Publishing Company, Incorporated, 1st edition, 2008
[10] M. J. Bauer, A. J. Cox, and G. Roson, Lightweight BWT construction for very large string collections, CPM 2011, volume 6661 of LNCS, Springer, 2011,pages 219–231
[11] Jens Stoye, Dan Gus.eld, Simple and flexible detection of contiguous repeats using a suffix tree. Department of Computer Science, University of California, 2014,vol. 270,issue 1-2,843-850
[12] Bonham-Carter O, Steele J, Bastola D, Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis,vol.6,2014,pages 890-905
[13] Ela Hunt Malcolm ,P. Atkinson, Robert W. Irving ,A Database Index to Large Biological Sequences, International Journalon VLDB,2001, pages 139-148
[14] Zhenqiang Tan Xia Cao Beng Chin Ooi Anthony K. H. Tung, The ed-tree: An Index for Large DNA Sequence Databases, Department of Computer Science National University of Singapore, 2003