International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4s2, December 2019

Fast and Accurate Identification of Short Tandem Repeats (STRs) Using Hash Function in DNA Sequences

S. Jawahar, P. Sumathi

Abstract: ***The main challenge in bioinformatics is the size and complexity of input datasets. Tandem repeats detection is important function in biology and medicine for phylogenic studies and diagnosing various diseases. Short Tandem Repeats (STRs) plays an important role in human genetic disease and for various regulatory mechanism and evolution. The mutation rate is higher in STR which leads to more biological research in this area. In our study at least two adjacent nucleotide patterns are considered as tandem repeats. The Short Tandem Repeats (STRs) is identified and investigated for diseases related mutation in human. The proposed algorithm Short Tandem Repeat using Hashing (STRH) uses hash table for fast storing and easy retrieval of values. The hash function generally hashes a longer string into much shorter string with fixed length. The analysis of STRH is made using five genes, HUMTH01, CSF1 Receptor, FIBRA, TPOX and VWF. Mostly the STRs in the five genes are tetranucleotide and contains perfect tandem repeat. The proposed STRH algorithm identifies more number of tandem repeats than the traditional algorithms.***

Keywords: ***Bioinformatics, tandem repeat, hash function, short tandem repeat (STRs), tetra nucleotide***

I. INTRODUCTION

Repetitive elements are found in coding and non-coding region, and also in intergenic regions of prokaryotes [5]. The repetitive elements in human are approximately 7% of the whole genome [6].

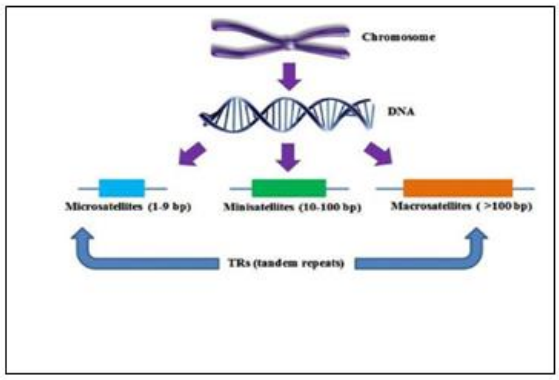

Figure 1: Different types of tandem repeats (TRs)

Manuscript published on 30 December 2019.
***Correspondence Author(s)**
S. Jawahar, Research Scholar, PG & Research Department of Computer Science, Government Arts College, Coimbatore (Tamilnadu), India.
Dr. P. Sumathi, Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore (Tamilnadu), India.

The repetitive sequence can be categorized into many types of groups namely, 1. Interspersed repeats,2. Tandem repeats, 3. Alu repeats, 4. Medium reiteration repeats. The interspersed repeats are scattered throughout the genome of human DNA which are either active or inactive copies of transposable elements.

There are two subcategories in interspersed repeats namely, 1.

Short Interspersed Nuclear Elements (SINE) and 2. Long Interspersed Nuclear Elements (LINE). The SINE accounts for upto 13% of human genome and LINE constitute upto 21% of human genome.

Tandem repeat sequences are also known as satellite DNA. The term satellite is used to refer DNA sequence which contains short repeats for specific motifs identification.

There are three subcategories in tandem repeat namely, 1. macrosatellites, 2. minisatellites and 3. microsatellites.

In macrosatellites sequences are located near centrosomes, telomeres and in Y chromosomes. The repeat length in this sequence is 171 base pair. The minisatellites sequences are also known as variable number of tandem repeats (VNTR) are located in non-coding region of DNA. The repeat length ranges between 9 base pair to 80 base pair.

The microsatellites are also known as short tandem repeats (STR) which consist of 1 to 9 base pair which can extend up to 150 base pair. According to the length of the repeat unit STRs are sub classified into mono, di , tri, tetra, penta and hexa nucleotide repeats. The total number of sub classification decreases as the size of the repeat unit increases. The dinucleotide repeats are most common STRs in human genome [3].

There are three subcategories of microsatellites namely, 1. Perfect, 2. Imperfect and 3. Compound. Perfect satellites are the identical repetitive units which contains only one repetitive unit. Imperfect satellites are small mutation units which may have been caused by insertion or deletion.

These satellites contain only one repetitive unit. Compound satellites contain perfect or imperfect sequences with two or more repetitions arranged successively with or without nucleotides between the bases [4].

Table 1: Commonly used Short Tandem Repeats (STRs)

Microsatellite categories	Description	Key Features	Example
Perfect	Exact	Identical Copies	ATC ATC ATC ATC ATC
Imperfect	Approximate	Substitution,insertion & Deletion (Mismatch)	ATC ATC AGC ATC AC
Compound	Fuzzy	Substitution, Multiple motifs	AT AT AT AT

Challenges of the problem

There are various major challenges in repetitive sequences,

1. Mining long sequence patterns in DNA.
2. Selecting the mined patterns efficiently.
3. Fault tolerance for interesting or repeat patterns.

The important operation in many applications is searching dataset and how fast data can be searched. The searching algorithms can be categorised into many types namely binary search, linear search, tree search, genetic algorithm etc. Hashing [2] is used for sorting and indexing information stored in hash table. It is used for finding specific object from group of objects. Hashing is mainly used to reduce disk space and access time for inserting and retrieving information. It can be implemented in two ways:

1. Convert input data into integer by using hash function and

2. Information can be retrieved easily using hash key.

Hash = hashfunc(key)

Index = hash%array_size

Hash function [1] aims to map dataset with large variable size to small and fixed size.

II. CONCEPTS AND DEFINITIONS

Definition 1: The hash table contains two columns: Input value (P) and replace value. The table used to store replace value is called replace table denoted by RT. The P contains single alphabet {a,t,g,c} and replace value contains integer number.

Table 2: Replace Value Table for DNA sequence

Input Value (P)	Replace Value
a	00
c	01
g	10
t	11

Hashing is a method used to store and retrieve the information from database or hash table. In this method the strings can be transformed into fixed length value called key. By using this method the information can be located and retrieved from the table efficiently.

There are various benefits namely, 1. overhead is reduced, 2. Locating information fastly, 3. Better optimised searching of information and 4. Ease of transfer. The information are sorted and indexed by using hashing methods. The hash table is used to store hashed values in system memory using hashing algorithm. Hash function is a process which transforms large size data into smaller fixed size data.

Hash table

Hash table is also called as hash map or scatter tables. This table uses hash function to map identifiers or keys to corresponding values. This table consists group of array for storing and accessing data using hash index. The short tandem repeat information is stored in the table. Hash table is allocated dynamically for storing each short tandem repeat.

Hash Function

A hash function contains group of characters which is called as key. This key is used to map certain value called as hash value. The hash value is normally smaller value than the original string of characters. This function locates and indexes items in database easier to find the shorter hash value than longer string value. There are three ways for identifying good hash function namely, 1. Perfect hash function, 2. Desirable hash function and 3. Trade-off function.

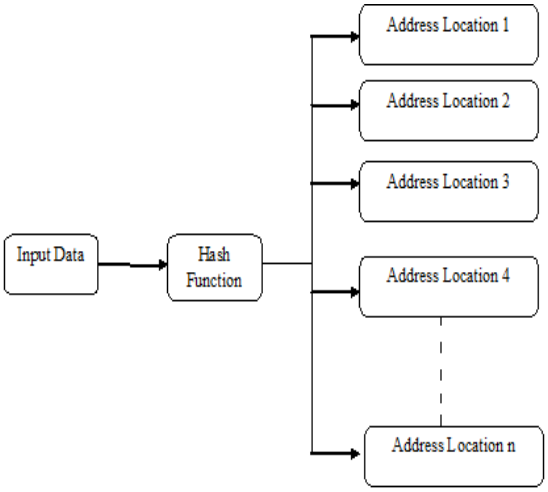

Figure 2: Hash functions and address locations

Algorithm: Short Tandem Repeat using hashing (STRH)

Step 1: *Input a DNA sequence of length n. { S = S1,S2,S3…..... ,Sn }*

Step 2: *A hash tables namely hash1 is initialised dynamically for storing integer values.*

Step 3: *The STR() scans the first sequence for identifying starting point of STR*

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Step 4: *STR() checks each position from starting point to the leftmost position of the sequence.*

Step 5: *The STRFilter() checks for the previous identified STR and avoids duplication*

Step 6: *The number of times the STR repeat is calculated.*

III. IMPLEMENTATION

A repeat region contains only repeating sequences. In DNA sequence usually has multiple repeats. The input is multiple DNA sequence which contains four characters {a,t,g,c}. The problem is to find the repeat patterns which are repeated at least thrice in mined pattern. The experiments on various genes are carried out on computer Intel Core i5, 3.7 GHz processor with 4GB RAM and Windows 7 operating system.

The program is written in java and net beans are used to run it. The input DNA sequence is stored in text file in FASTA format and program reads the input file to find all STR patterns.

The data sets used are real life DNA sequences which are downloaded from the website of National Center for Biotechnology Information (NCBI) [7]. By using the advanced search technique in NCBI the required genes can be downloaded as,

The following parameters are used to download DNA data set using advanced search technique,

(a)Search as Category=”Nucleotide”, (b)Organism =”human” and (c) gene=”Gene name” (d) All Fields=” Related gene information”.

Table 3: Short Tandem Repeat in various genes

Sequence ID	Gene Name	Category(Perfect/Imperfect)	Short Tandem Repeat (STR)
1	HUMTH01	perfect	TCAT GTAA
2	CSF1 Receptor	perfect	AGAT
3	FIBRA	perfect	TTTC TTTT
4	TPOX	perfect	AATG
5	VWF	Imperfect	TCTA TCTG

The table represents short tandem repeat for various genes namely, 1. HUMTH01, 2. CSF1 Receptor, 3. FIBRA, 4. TPOX and 5. VWF respectively. The category represents two possibilities either perfect STRs or Imperfect STRs. The genes HUMTH01, CSF1 Receptor, FIBRA and TPOX contains perfect STRs. These four genes STRs are tetranucleotide which contains 4bp in sequence. The VWF gene contains imperfect STR ie., TCTA followed by TCTG. This gene STR is also tetranucleotide which contains 4bp in sequence.

IV. RESULTS AND DISCUSSION

In this section the proposed algorithm is compared with Tandem Repeat Finder and Repeat Master. In proposed algorithm hash function and hash table is used for fast access of the DNA sequence. By using this method indexing and sorting of the sequence are more efficiently handled.

Table 4: Comparison of STRH algorithm with Tandem Repeat Finder and Repeat Master

S.No	Gene Name	STR using Tandem Repeat Finder	STR using Repeat Master	Total Number of STR using STRH
1	HUMTH01	5	7	9
2	CSF1 RECEPTOR	6	10	12
3	FIBRA	10	16	21
4	TPOX	9	9	11
5	VWF	16	19	20

The total no. of Short Tandem Repeat using STRH algorithm is higher than Tandem Repeat Finder and Repeat Master.

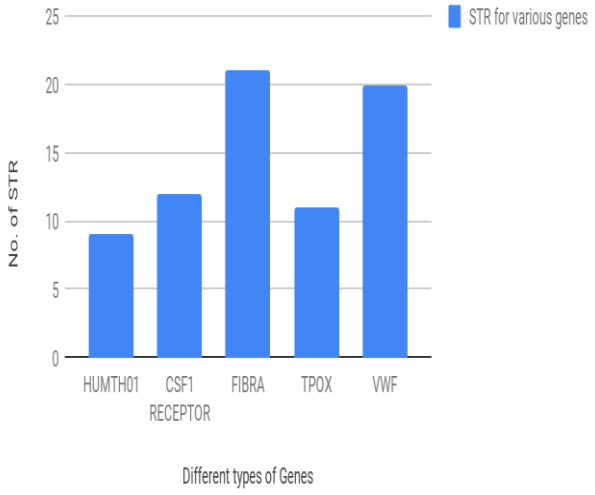

Figure 3: Different genes vs No. of STRs

The bar graph in figure 3 represents the relationship between different types of genes (5 genes) and number of short tandem repeats identified using the proposed algorithm. The FIBRA and VWF gene almost has same number of short tandem repeats.

HUMTH01

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

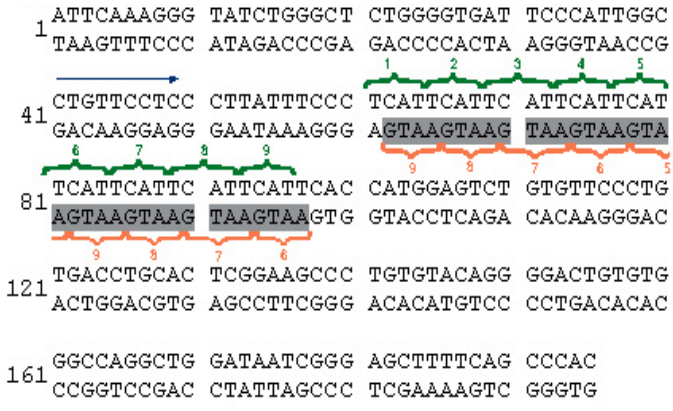

Figure 4: STR in HUMTH01 gene

The HUMTH01 is Human tyrosine hydroxylase gene which is located in chromosome 11. The common STR in this gene is TCAT and STR length is 9. The STRs identified using Tandem Repeat Finder is 5 and Repeat Master is 7 in length.

CSF1 Receptor

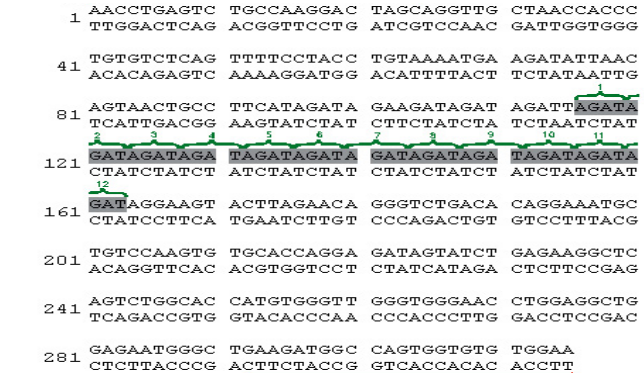

Figure 5: STR in CSF1 Receptor gene

The CSF1 Receptor gene is also known as macrophage colony-stimulating factor receptor (M-CSFR) which is located in chromosome 5. The total length of chromosome is 60081bp. The common STR in this gene is AGAT and STR length is 12. The STRs identified using Tandem Repeat Finder is 6 and Repeat Master is 10 in length

FIBRA

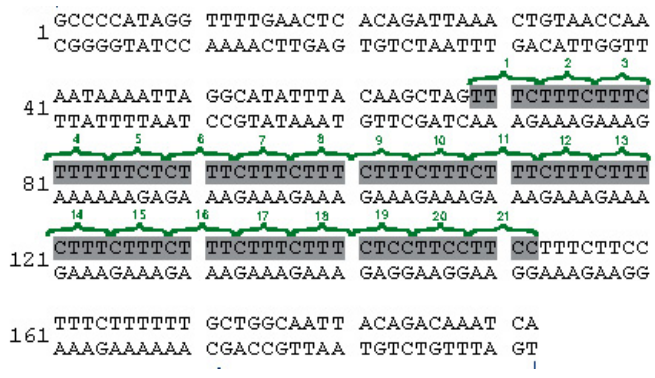

Figure 6: STR in FIBRA gene

The FIBRA gene is also known as fibrinogen gene which is located in chromosome 4. The total length of chromosome is 7576 bp. The common STR in this gene is TTTC and STR length is 21. The STRs identified using Tandem Repeat Finder is 10 and Repeat Master is 16 in length.

TPOX

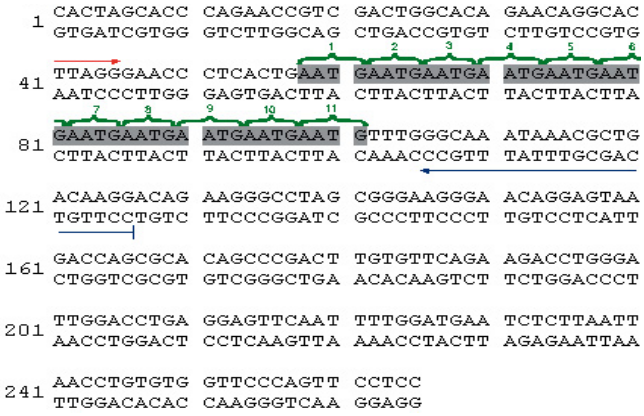

Figure 7: STR in TPOX gene

The TPOX is human thyroid peroxidase gene which is located in chromosome 2. This gene is also known as HUMANTPOX. It contains 17 exons encoding a protein 933 amino acids. The common STR in this gene is AATG and STR length is 11. The STRs identified using Tandem Repeat Finder is 9 and Repeat Master is 9 in length

VWF

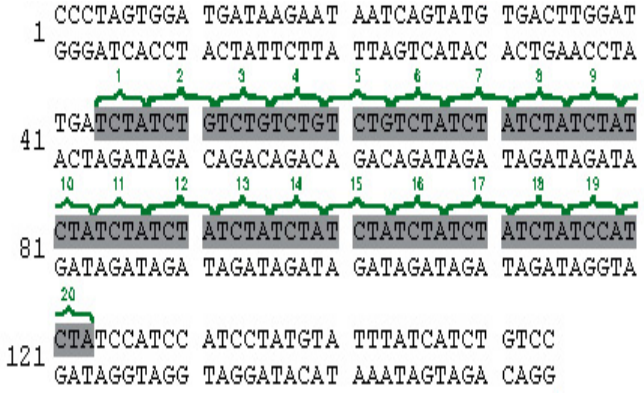

Figure 8: STR in VWF gene

The VWF is Von Willebrand Factor gene which is located in chromosome 12. This total length of VWF gene is 175896bp. It contains totally 52 exons with type1, type2 and type3 inherits. The common STR in this gene is AATG and STR length is 20. The STRs identified using Tandem Repeat Finder is 16 and Repeat Master is 19 in length.

V. CONCLUSION

In this paper the proposed Short Tandem Repeat using Hashing (STRH) algorithm is used for identifying tandem repeats in DNA sequences. The search accuracy and number of tandem repeat identification is efficient in the proposed algorithm. Totally five genes are used for STR process and mostly the repeats are perfect. The STR in the five genes are tetra nucleotide which contains four amino acids in pattern. Compared to the traditional algorithms STRH identifies more STR efficiently and accurately. In future the overall performance of the algorithm can be enhanced by using the collision avoidance mechanism in hash table and various other filters for the unwanted short tandem repeats.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

REFERENCES

1. Carter, J. Lawrence, and Mark N. Wegman. "Universal classes of hash functions. Proceedings of the ninth annual ACM symposium on Theory of computing. ACM, 1977.
2. Dietzfelbinger, Martin, et al. "Dynamic perfect hashing: Upper and lower bounds." SIAM Journal on Computing 23.4 (1994): 738-761.
3. International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921
4. Subramanian S. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003;4:R13
5. S. Leclercq, E. Rivals, and P. Jarne, “Detecting microsatellites within genomes: significant variation among algorithms,” BMC Bioinformatics, vol. 8, article 125, 2007.
6. K. G. Lim, C. K. Kwoh, L. Y. Hsu, and A. Wirawan, “Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance,” Briefings in Bioinformatics, vol. 14, no. 1, Article ID bbs023, pp. 67–81, 2013.
7. National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)