IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 1, 2021, accepted November 30, 2021, date of publication December 10, 2021,
date of current version December 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3134284

An FPGA-Based Accelerated Mutation Detection System for the Tumor Suppressor Gene

MUHAMMAD IBRAHIM[1], OMER MUJAHID[2], NAJIB UR REHMAN[1], AZHAR QAZI[1], ZAHID ULLAH[3], AND TAMA FOUZDER[4]

[1]Department of Electrical Engineering, CECOS University, Peshawar 25000, Pakistan
[2]MICE Lab, Department of Electrical, Electronics and Automation Engineering, University of Girona, 17004 Girona, Spain
[3]Department of Electrical and Computer Engineering, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur 22650, Pakistan
[4]Department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh (ULAB), Dhaka 1209, Bangladesh

Corresponding author: Zahid Ullah (zahid.ullah@fecid.paf-iast.edu.pk)

ABSTRACT This paper proposes a novel fast mutation detection system that looks for mutations in the tumor suppressor gene, also known as TP53. Mutations are modifications in the nucleotide sequences of the human genome and may be caused by various factors, such as exposure to radiation, sunlight, smoking and replication errors. Mutations in TP53 are the most common cause of cancer and early detection may prevent cancer from happening. The proposed system utilizes the high matching speed of a logic-based content-addressable memory (CAM) for mutation detection along with a hamming distance calculator that specifies the exact location of the mutation. The proposed system implementation is carried on a Xilinx Virtex®-7 field-programmable gate array (FPGA) and demonstrates a low match time of 0.18 μs, which is much faster compared to the state-of-the-art systems.

INDEX TERMS CAM, FPGA, SR-based CAM, gene mutation detection, TP53 gene, p53, tumor suppressor gene.

I. INTRODUCTION

The building block in every human body is known as a cell that performs different tasks. A combination of cells make the human body organs [1]. Each cell contains deoxyribonucleic acid (DNA); mostly located in the nucleus of the cells or rarely, in mitochondria. The information required to build and maintain an organism (known as an organism's genetic blueprint) is encoded by the DNA [2]. Entire human DNA comprises of around 3 billion bases [3], and above 99% of those bases are same in all the human beings. The DNA composes of four chemical bases namely adenine (A), cytosine (C), guanine (G), and thymine (T). Two strands pair up together with the help of a bond; the base pairs are A with T and C with G [4].

Nucleotide refers to the combination of each DNA base with a sugar molecule and phosphate. A spiral ladder-type shape is formed from a nucleotide which is known as double helix [5]. Genes are the small sections taken from the long chain of DNA. They are the essential unit of heredity (both functionally and physically) and in people, the number of DNA bases varies in each gene from a few hundred to over two million. According to the human genome project approximation, humans have approximately 20,000 to 25,000 genes [6]. A mix of various genes together create chromosomes. Children acquire two sets of chromosomes; one from each parent. That is the reason why they have two copies of every gene. There are 23 pairs of chromosomes in human beings [7].

Genes contain the information regarding the formation of the molecules called proteins. It plays a vital role in cell functionality. Each cell depends on these proteins; present in thousands of number in order to perform the right task at the right time [8].

Mutated versions of genes known as alleles are created when a small-scale variation occurs in DNA bases' gene sequence. These variations in gene sequence can prevent the proteins from working properly [9]–[11]. Inability to perform the tasks can lead to a change in a cell or an organ and lead to disease. Thus, a variation of a gene can cause disease rather than the gene itself. Assume that if someone has ''the cystic fibrosis'' [12], the condition is caused by a variation of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. People without cystic fibrosis have a copy of the CFTR gene as well [13].

The most mutated gene in human cancer is TP53. The TP53 gene is known as a tumor suppressor; it provides instructions for making tumor protein p53. It is located on chromosome

The associate editor coordinating the review of this manuscript and approving it for publication was Zhen Ren.

17 in humans and controls the cell division by keeping the cells from proliferating too fast. More than 50% of human cancers contain mutated TP53 gene [14], therefore it plays a vital role in cancer prevention [15]. The aim of the proposed work is to use content-addressable memory (CAM) for rapid mutation detection in TP53 gene. The proposed system displays high speed and throughput as compared to software-based systems [16], [17].

The key contributions of this research work include:

- To the best of our knowledge, this is the first time a mutation detection system has been implemented using CAM on an FPGA device.
- This mutation detection system performs better than the state-of-the-art systems in terms of power efficiency.
- Along with very fast matching speed, our mutation detection system also specifies the exact location at which the mutation has occurred.

The rest of the paper is organized as: background study and related work are discussed in section II and III. Section IV focuses on computer memory and its types. Using logic-based or slice register (SR)-based CAM architecture for mutation detection in human genome is discussed in detail in section V. The proposed system architecture and FPGA implementation are discussed in section VI. While section VII concludes this paper, having results, comparison and future work.

II. MUTATION IN HUMAN GENOME

The mutation is a change happening in DNA sequence, either by ecological factors like daylight, radiations, smoking, and so forth, or unintentionally during DNA replication [18]. These transformations in the human genome are otherwise called variations [19]. The human body comprises of 23 sets of chromosomes comprised of a long chain of DNA, which is practically equivalent to 3.2×10^9 base sets or 20-25k genes [20]. There are two subcategories in which each set of 23 chromosomes can be described; autosomes and sex chromosome, which are 1-22 and 23^{rd} respectively [21]. Various kinds of changes can occur in these strands, going from revamping of the chromosome or duplication, or even a single nucleotide polymorphism (SNP) to the deletion of nucleotide [21]. Genetic diseases can be divided into two principle classes, i.e, chromosomal and mendelian (inherited) disorder.

A. CHROMOSOMAL DISORDERS

Typical vegetal or somatic cells go through a division cycle known as mitosis to accelerate their development, whereas meiosis division occurs in germ cells. Process reduction division (Meiosis I) and subsequently equational division (Meiosis II) are carried out during the meiosis. The development of a primary egg cell (or ovum) into a mature ovum, known as oogenesis, begins in the female embryo in the twelfth week, but it ends around the twentieth week at the phase of meiosis I, once the chromosomes have combined to form bivalents, quadruplicates, & replicate [22].

Misprints or inconsistencies in chromosome pairing, structural damage, and crossing-over are all causes of chromosomal disorders. Aneuploidy occurs when a cell has an abnormal number of chromosomes. A human cell, for example, may have 45 or 47 chromosomes instead of the usual 46. It is caused by abnormalities in the division of chromosomes during meiosis I and II [23]. Duplication, deletion, reciprocal T, translocation, Robertsonian T, and so on are included in chromosomal variations in human disorders. Gene-based mutation human disorders can be categorized as duplication, point mutation, deletion, splice mutation, insertion mutation, and dynamic mutation [21].

On the basis of functionality, the mutations can be subcategorized as: *Gain of function mutation* & *Loss of function mutation*. An increase in gene products or activity is caused by the gain of function mutation, which, in various cases, results in the creation of toxic products which lead to diseases like cancer, etc. On the other hand, the loss of function mutation results in the loss or decrease of gene products or activity.

B. MENDELIAN DISORDER

In light of some standards founded by Gregor Mendel in the nineteenth century, the patterns of inheritance characteristics or heredity are determined [24]. These principles include; autosomal dominant (AD), autosomal recessive (AR), X-linked recessive (XR), X-linked dominant and Y-linked disorders [21].

Moreover, some subjects having disease-causing mutations do not exhibit any signs or symptoms of the disorder, while others do. For instance, a mutation in the Breast Cancer type1 (BRCA1) gene may cause breast cancer in certain individuals while it probably will not be formed into the tumor in others [25].

Although certain genetic disorders do not follow any of the above-mentioned inheritance patterns. Mitochondrial illnesses including trinucleotide expansion abnormalities and genomic imprinting problems [26] are examples of non-Mendelian patterns of inheritance.

III. PRIOR WORK ON MUTATION DETECTION

This section of the study summarizes the most recent prior research in the field. As discussed in section II, change occurring in DNA sequence can cause several genetic disorders. Noval detection techniques are required to identify these mutations. Mutation detection has a vital role in genetic diagnosis including confirmational diagnosis, presymptomatic testing, and identity/forensic testing, etc.

In order to discuss the recent work done in the field of mutation detection, prior work can be categorized into two sub-categories, i.e, software-based and hardware-based systems. In software-based algorithms, [16] presented three algorithms: First-Last Pattern Matching (FLPM) algorithm, Processor-Aware Pattern Matching (PAPM) algorithm and Least Frequency Pattern Matching (LFPM) algorithm. FLPM algorithm performs character-based comparison, in which the

first and last indexed character of the query string is compared with the stored pattern. The query string window comprises of six characters. In order to reduce the window size and to improve run-time, word-based comparison (a word comprised of few characters) approach is used in PAPM and LFPM algorithm.

LFPM [16] surpassed the other two algorithms, in term of time required for matching and reducing query pattern window size. As LFPM only look for the word of query string in the stored string that has the lowest occurrence frequency. The main disadvantage of these three algorithms is that it do not perform parallel operation. Secondly, the shifting of query pattern window over a particular span of stored pattern also requires more memory.

Discrete to Continuous (DTC) algorithm is developed and presented in [17], in which superposition of a discrete test points over continuous reference points is performed. This algorithm is also based on character matching approach. The pre-processing of searching the superposition of test pattern in reference pattern is very time consuming. The nucleotides of test pattern and reference pattern is represented by an abscissa and an ordinate in a space of 2 dimensions (2D). The issue with this algorithm is that the distance calculation between the test pattern and reference pattern is done by a non-linear equation (having a square root), which makes its hardware implementation nearly impossible because it exploits the hardware resources.

The hardware-based system (for mutation detection) presented in [27] is using approximate string matching (ASM) approach. The hardware is named as genome approximate string matching (GenASM). The main disadvantage of GenASM is that it is a random access memory (RAM) based system, which performs sequential computation and thus it lacks in parallel search capability.

Therefore, a question arises, is there any way to tackle the issues in the aforementioned systems? The answer is yes. The proposed system uses CAM which has the capability to compare the test pattern concurrently with the reference; thus reducing the execution time. Furthermore, the system presented in this paper is also resource efficient. Architecture and methodology of the proposed system are explained in detail in section IV, V and VI.

IV. COMPUTER MEMORY & CLASSIFICATION

Computer memories are designed to consume less power or read/write data at a high speed, depending on discrete requirements or parameters. Additionally, on the basis of data storage i.e. permanent or temporarily depending on the power, computer memories can be categorized as volatile and non-volatile [28]–[30].

These days, several type of memories exist; however, primary memory and secondary memory are the basic type of memories. System memory is actually the primary memory while storage is referred to as secondary memory. Memory is further categorized as CAM, read-only memory (ROM), and RAM. As the proposed system uses CAM; therefore

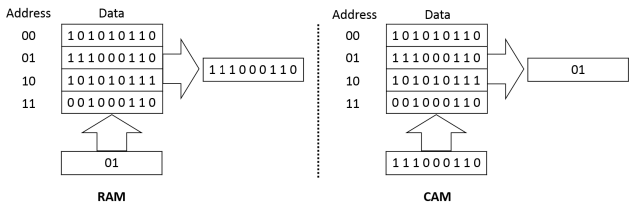


FIGURE 1. Difference between RAM and CAM.

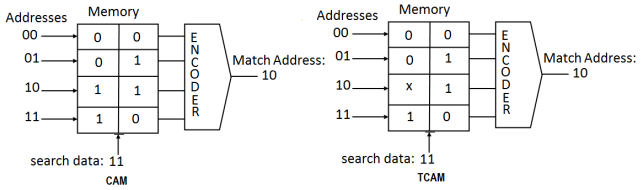


FIGURE 2. Classification of CAM (i) BiCAM and (ii) TCAM.

section IV-A focuses on CAM, its types, and how it differs from RAM.

A. CAM, ITS TYPES, & COMPARISON WITH RAM

In contrast to RAM, which takes an address as an input and outputs stored data, a CAM's input takes data and the output gives the stored data addresses as shown in Fig. 1. Due to parallel processing being carried out in CAM, its speed is faster than RAM. On the basis of functionality, CAM is divided into two types as stated below;

1) Classical Content Addressable Memory (CAM): In CAM, every CAM cell typically contains a storage circuit and a comparator. When the search data is passed to the CAM cell, the entire memory is simultaneously searched, and the output with the highest priority is selected using an encoder. Fig. 2 illustrates the two types of traditional CAM, namely (a) *Ternary Content Addressable Memory (TCAM)* and (b) *Binary Content Addressable Memory (BiCAM)*, based on the operation mechanism. For exact matching, "*BiCAM*" is used since it can only store "0" or "1" but "*TCAM*" is used for partial matching because it can store a third state, which is symbolised by the letter "X" and is known as the "don't-care bit". The mask bit in *TCAM* is represented by the 'X' bit; an exact match will occur if the mask bit value is '0', similar to a *BiCAM*, but if the mask bit value is '1', the match will always appear as 'true', regardless of the search query supplied. CAM is now used in a variety of applications, including pattern matching, data compression, packet categorization, and image processing [31]–[34].
2) Field Programmable Gate Array (FPGA) based CAM: In the 1980s, a programmable arrangement recognized as Programmable gate array (PGA) or Programmable logic array (PLA) later on called Programmable logic devices (PLDs) were introduced with some programming and employing AND-OR structures. FPGA,

which was initially conceived by Xilinx, is created by combining programmable ROM (PROM) with PLD [35], [36]. The FPGA is built in such a way that it may be reprogrammed by the designer after it has been manufactured. FPGAs are frequently configured using the hardware description language (HDL). It is used for developing and testing digital circuits [37]. Subsequently, CAM is an expensive memory as compared to RAM, so its functionality can be emulated using FPGA hardware. Remember, there is no built-in CAM input port on FPGA though we can use logical resources available on the FPGA to emulate CAM functionality. Static RAM (SRAM)-based CAM, logic (flip-flops)-based CAM, slice register (SR)-based CAM and Look Up Table (LUT)-based CAM are the types of FPGA-based CAM.

V. METHODOLOGY

The requirement of large memory units for huge biological data and high processing time are the issues faced by Next Generation Sequencing (NGS) techniques [38], [39]. In order to address the processing time issue, CAM is used for the detection of mutations in human genome and a single nucleotide polymorphism (SNP) in DNA sequence is targeted in this research work. Accuracy, expensive fast computation and memory requirement, are the issues faced by already developed SNPs detection algorithms [40]. Therefore, a fast, accurate and scalable mutation detection system is proposed. The proposed system is much speedy than RAM based system due to the concurrent search/match mechanism of CAM; where the exact matching is achieved by using SR-based or logic-based CAM architecture [32].

The TP53 gene sequence is taken into consideration in the proposed system to detect mutations, as it mutates in most types of human cancers [41]–[44]. The TP53 gene codes or provides instruction for making a tumor protein known as p53. It functions as tumor suppressor protein and controls the proliferating or an uncontrolled cell division process. The protein p53 is altered in most cancers; therefore, a well-timed diagnosis of the mutated TP53 gene is essential.

Moreover, to recognize the number of nucleotides mutated in the particular section of the genome, a hamming distance calculating algorithm is used. Where the test pattern is compared with the benchmark datasets. Point mutation is a type of mutation where a single nucleotide is embedded, changed, or erased from the DNA chain. This added feature of incorporating the hamming distance calculator aids in detecting point mutations and determining the scale of mutation.

A. THE GUARDIAN OF THE GENOME, p53

TP53 is the gene that codes for the tumor suppressor protein ''p53''. Every cell in the human body contains this tumor suppressor protein. It guards against any kind of damage to DNA. Poisonous compounds, smoking, destructive radiation, and bright ultraviolet (UV) beams are just some of the things that can harm DNA. The p53 continuously checks for the damaged DNA as it is directly tied to the DNA chain. If any damaged part is found in the DNA chain, p53 assumes a fundamental part in fixing it. It either informs the respective cell to undergo self-destruction/eliminating the damaged part or try to repair it. The p53 activates other genes to fix the damaged part of the DNA if it is repairable else it stops the cells from cell division (mitosis). Accordingly, it forestalls the development of tumors.

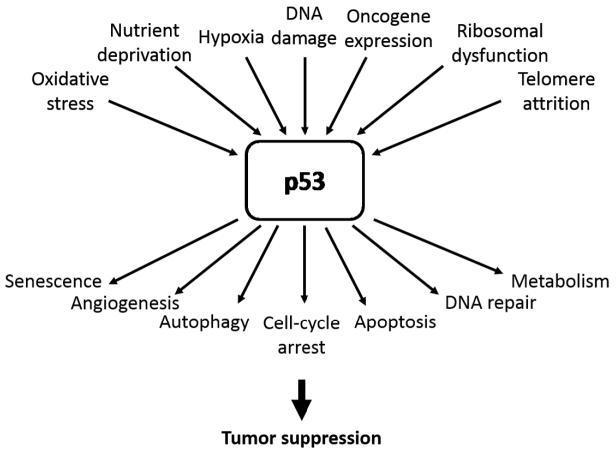


FIGURE 3. Usefulness of the guardian of the genome (p53).

The tumor protein p53 is also known as ''the guardian of the genome'', as it monitors the DNA chains, controls the mitosis process and can repair the genome if required. It is crucial for keeping the human genome safe from mutations. Fig. 3 depicts the entire utility of p53. The p53, as shown in Fig. 3, deals with the issues outlined by the arrows pointing toward it, as well as the strategies illustrated by the arrows pointing away from it. All these functionalities of the p53 contribute towards tumor suppression.

Freely accessible benchmark datasets of the national center for biotechnology information (NCBI) and the international agency for research on cancer (IARC) have been used to acquire the TP53 gene or p53 protein succession [45]. The combination of three nucleotides results in the formation of a codon. There are 394 codons in the p53 genomic sequence. In the revised reference sequence, codon position 72 CCC (pro) is mentioned, whereas CGC (arg) is indicated here [45]. This data (genome sequence) is then mapped into an SR-based CAM.

B. ARCHITECTURE OF SR-BASED CAM

In this research work, we employed SR-based BiCAM to achieve quick and precise genomic pattern matching. The proposed CAM simulates the functionality of a logic-based higher performance binary CAM (LH-CAM) [32], but instead of sending matched addresses to the output [32], the address of the mismatch will be provided by the CAM in our system. The SR-based CAM used in our system is shown in Fig. 4.

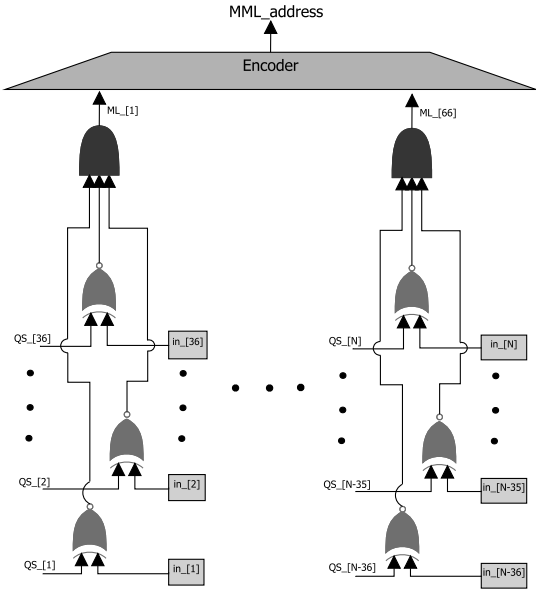


FIGURE 4. SR-based CAM architecture (ML: match-line, QS: query string, in: stored string, MML_address: mismatch-line address).

TABLE 1. DNA bases representation in binary.

Nitrogenous Bases	Adenine (A)	Cytosine (C)	Guanine (G)	Thymine (T)
Representation in Binary	00	01	10	11

It is comprised of a two-dimensional array of memory cells known as memory units (MUs). Each MU has a 1-bit storage and a 1-bit comparator, much like a traditional BiCAM cell. All of the MUs in a row are connected by a single line known as the match-line (ML). ''M'' stands for the total number of bits in each CAM word, or all the MUs in each ML are combined into an (M-bit) vector. Thus, a BiCAM with a dimension of 512 × 36 will have 512 MLs, with each ML containing 36 MUs. The encoder is also linked to all of the MLs. The mismatch address (MML_address) of the ML where the mismatch occurred will be provided at the encoder's output. A single mismatch in the MU of a certain ML causes a mismatch in that match-line, much like in conventional BiCAM.

C. THE p53 GENOMIC SEQUENCE MAPPING

Since DNA has four bases in total, as shown in Fig. 5, we could use Eq 1 to encode individual bases with two binary bits.

$$\text{Total number of bases} = b = 2^m \tag{1}$$

''b'' denotes the total number of nucleotides, and ''m'' denotes the number of bits per nucleotide. If each base is represented by two bits, $m = 2$, then $b = 2^2 = 4$. Two flip-flops (FFs) or MUs are required in the SR-based CAM to map a single base or nucleotide. The binary representation of each base is shown in Table 1.

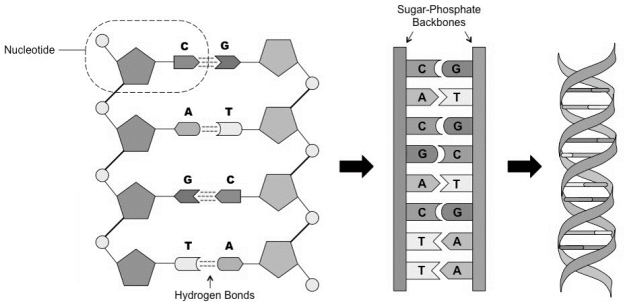


FIGURE 5. Nucleobases or nitrogenous bases of DNA.

TABLE 2. SR-based CAM sequence mapping for the "p53."

Address	Match-lines (MLs)	p53 Sequence Codon (Cs)		Binary Representation Vectors (Vs)	
0	ML_0	C_1	ATG	V_1	001110
0	ML_0	C_2	GAG	V_1	100010
0	ML_0	C_3	GAG	V_1	100010
0	ML_0	C_4	CCG	V_1	010110
0	ML_0	C_5	CAG	V_1	010010
0	ML_0	C_6	TCA	V_1	110100
1	ML_1	C_7	GAT	V_2	100011
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
65	ML_{65}	C_{394}	TGA	V_{66}	111000

A codon is a sequence of three nucleotides. The p53 sequence is made up of 394 codons [45], totaling 1,182 nucleotides. Since each nucleotide is represented by two binary bits, the whole sequence of p53 requires a total of 2,364 bits to map.

Accordingly, the CAM is 66 × 36, with 66 horizontal rows (MLs) and each row consisting of 36 bits or 36 MUs. Hence, it has 66 vectors (Vs), each of which is 36 bits long. The mapping of the TP53 gene sequence to the SR-based CAM is shown in Algorithm 1. The codons are distributed as 6 codons per ML to maximize resource use. Every horizontal ML has 18 nucleotides, requiring 36 bits per ML to map them. Table 2 shows a mapping example using Algorithm 1. The mapping difficulty is O(1), although in prior CAM architectures, the updating difficulty increased exponentially with the depth of the CAM and also required pre-processing [32]. Consequently, SR-based CAM is easy and quick to update.

Algorithm 1 Mapping Algorithm

1: Input : p53 Gene Sequence
2: Output: Condons Ensemble
3: Number of Condons (C_s) per match-line: $n = 6$
4: **for** $i \leftarrow 0$ to 65 **do**
5: Map n Codons Ensemble to Vi of the SR-based CAM
6: **end for**

D. THE MATCHING MECHANISM

Each MU has the ability to store and compare a single bit. The test pattern or query string (QS) is compared concurrently

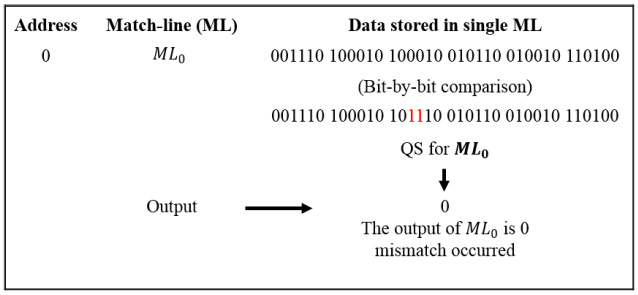


FIGURE 6. Bit-by-bit comparison process in each match-line.

with all the vectors (Vs) using a bit-by-bit comparison procedure. The matching/searching difficulty is also O(1) due to concurrent processing.

As every MU has a single bit comparator that compares the QS to the stored Vs bit by bit, if a single bit mismatch happens in any ML, the entire ML turns low (logic 0), however if the bits stored in Vs match the input QS, the entire ML remains high (logic 1). The mismatch-line address (MMA) is delivered to the output if there is a mismatch; else, default address (DA) is delivered to the output. 6 codons or 18 nucleotides are available in the first match-line (ML_0), for example. Table 2 shows its binary representation, and Fig. 6 illustrates the matching procedure.

Due to the 2nd nucleotide in the 3rd codon, a mismatch occurred in ML_0. As a result, the entire match-line becomes low (logical 0), and MMA is passed to the output (in this example the mismatch-line address is 0).

Algorithm 2 Concurrent Search Algorithm

```
Input : 36-bit Query string (QS)
Output: Mismatch-line address (MMA) or default address (DA)
Search vectors Vi concurrently, i = 1, 2, 3 . . . , 66
if QS matches Vi then
    Match occurs and DA is sent to output
else
    Mismatch occurs and MMA is sent to output
    Respective QS ⊕ Vi stored at MMA
    Number of mutated nucleotide extracted
end if
```

As illustrated in Fig. 7, the proposed system is made up of three primary blocks; SR-based CAM and matching/mismatching have already been explored extensively in the preceding paragraphs. The test sequence is fed into SR-based CAM, which employs Algorithm 2 to detect the mutation by comparing it to the benchmark stored sequence. As per the architecture of SR-based CAM, each MU has 1-bit storage and a 1-bit comparator. Thus, the search vector is compared bit-by-bit with the stored vector at every ML. If a single bit mismatch occurs, the entire ML will turn to a low state and the system will indicate that a mutation has been found. The mutated pattern is then given to the hamming

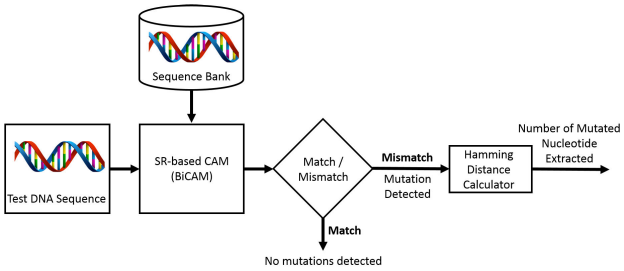


FIGURE 7. The proposed mutation detection system's block diagram.

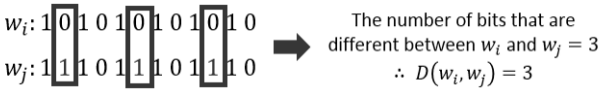


FIGURE 8. An example of calculating hamming distance.

distance calculator, which calculates the number of altered nucleotides. If there has been no mutation found, the system will return the default value, DA.

E. HAMMING DISTANCE CALCULATOR

The hamming distance calculator compares the reference benchmark pattern to the mutated or test pattern. As shown in Fig. 8, the number of locations where the QS differs from the SP is the Hamming distance between two vectors (test pattern (QS) and stored pattern (SP), d(QS,SP).

In order to elaborate the functionality of the hamming distance calculator, let suppose a mismatch occurs at a particular ML. The test vector at that ML will then be provided to the hamming distance calculator and bit-by-bit comparison (XOR) is perform with the reference vector. The priority encoder will receive the outputs of all the XORs in order to determine the exact number of altered nucleotides. This added feature helps us determine the scale of the mutation.

VI. ARCHITECTURE & FPGA IMPLEMENTATION

For implementation, Verilog HDL is employed, while Xilinx's ISE Create Suite 14.5 is utilized for design purposes. The proposed system is implemented and tested on the Virtex-7 device XC7VX330T FPGA using the Xilinx simulation tool. For the analysis, post-place and route results are taken into account.

A. ARCHITECTURE OF PROPOSED MUTATION DETECTION SYSTEM

The FPGA-based mutation detection system is built with a 66×36 size SR-based CAM and encoding circuits. The depth of the proposed mutation detection system is 66, with 36 bits stored in each match-line vector. Logic gates, a comparator, and encoding circuitry make up the overall system.

The proposed system have four input ports (CLK, in, QS, and set) and one output port (ENCRes), as shown in Fig. 9. The clock is represented by "CLK", the reference benchmark sequence is represented by "in", the query string is

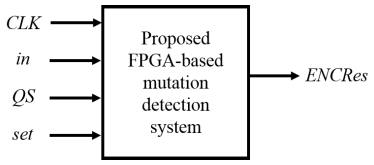


FIGURE 9. I/Os of the proposed system.

TABLE 3. Comparison of different systems based on detection time.

S. No	System Architecture	Design type	Size	Detection Time
1.	Proposed System	Hardware based	1188 (nucleotides)	0.00018 ms
2.	GenASM [27]	Hardware based	64bp – 320bp	0.00225 ms
3.	DCMP [16]	Software based	25 (nucleotides window)	3.5 ms
4.	BF [16]	Software based	25 (nucleotides window)	2.4041 ms
5.	BM [16]	Software based	25 (nucleotides window)	0.8966 ms
6.	FLPM [16]	Software based	25 (nucleotides window)	0.1852 ms
7.	PAPM [16]	Software based	25 (nucleotides window)	0.0024 ms
8.	LFPM [16]	Software based	25 (nucleotides window)	0.0018 ms
9.	DTC algorithm [17]	Software based	1000 (nucleotides)	17 ms

TABLE 4. Mutation detection speed of the proposed system.

System Architecture	Proposed System
Matching Type	Exact
Speed (MHz)	356.7
Time per matchi-line (ns)	2.803

represented by "QS", and "set" acts as a switch to activate a certain match-line. While "ENCRes" (encoder result) is the output port that displays the address of the mismatch. In addition to all of this, the hamming distance is computed and shown on the console.

VII. RESULTS AND PERFORMANCE COMPARISON

For performance comparison, based on the most relevant research work. A comparison is made between the proposed system and recent research [16], [17], [27].

A. SPEED OF THE PROPOSED SYSTEM

The CAM in proposed system is made up of SRs, LUTs, and logic gates. Furthermore, it can complete the matching operation in just two clock cycles. The proposed system performs better in terms of match time as compared with other mutation detection systems in literature. The detection time of various software and hardware-based systems are compared in Table 3.

The proposed system has 66 match-lines, and the matching process takes 2.803 nanoseconds per match-line. All of the processes are carried out simultaneously due to the usage of CAM, thus resulting in a system speed of 356.7 *MHz*. Table 4 provides the details.

B. POWER CONSUMPTION

According to the XPower Analyzer results, the presented technique consumes 27 mW of dynamic power. The power consumption of the GenASM system [27] built on FPGA is compared with the presented system. The GenASM hardware is made up of on-chip SRAMs, whereas the proposed system hardware is entirely made up of SRs and LUTs. They have

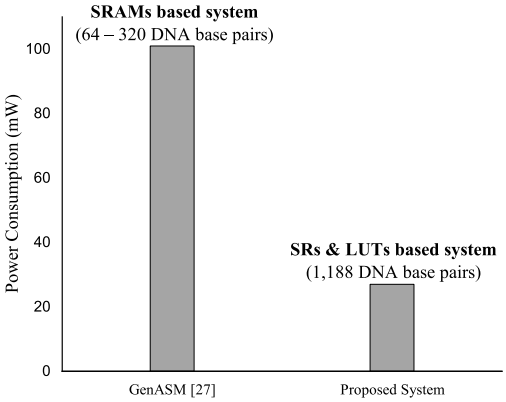


FIGURE 10. Power consumption of proposed system.

employed SRAMs to simulate parallel computing capabilities; however, the proposed system uses SR-based CAM to provide the same capability while also being faster and more power-efficient. According to the results in Fig. 10, it is 3.74 times more power-efficient than the GenASM. The GenASM uses 101 mW of power and has been tested with DNA base pairs ranging from 64 to 320. On the other hand, the proposed system has been tested on 1,188 DNA base pairs.

C. RESOURCES UTILIZATION

FPGA resources like Block RAMs, SRs, LUTs are used for the implementation of the proposed mutation detection system. Table 5 lists the resources used by the proposed system, as well as the resources that are available.

TABLE 5. Report on resource allocation.

Resources	SRs	LUTs
Utilized	2,449	2,452
Available	408,000	204,000
Utilization Percentage	1%	1%

Table 5 shows that the proposed system is resource efficient, since it uses just 1% of the SRs and LUTs. After the system's post-place and route implementation, the resource consumption report is generated. For rapid and efficient mutation detection, the presented system uses SR-based BiCAM, as previously stated. Since the CAM is 66 × 36, about 2,376 SRs are required for CAM implementation, with the remainder used for logic implementation (that includes XOR, etc).

VIII. CONCLUSION & FUTURE DIRECTIONS

Parallel processing, often known as parallel computation, is vital in the detection of mutations. The application of CAM to the detection of genomic mutations is a novel method. This approach will contribute to the development of a fast, power-efficient, accurate, and cost-effective mutation detection system for the entire genome in the future. In comparison to recently published work, the results indicate that the proposed system is power-efficient (consumes 27 mW), has

minimal execution time (0.18 μs), and is substantially faster (356.7 MHz). This can contribute in the early detection of various diseases, including cancer and viruses that adapt and evolve as they infect and proliferate in their hosts' cells. In the current pandemic situation of COVID'19, this system can help in identifying the mutation of coronavirus. Moreover, the proposed system is FPGA-based, which means that it can be reconfigured. In the future, we may be able to construct a genome approximate matching method or system.

REFERENCES

[1] U. Lucia and G. Grisolia, "How life works—A continuous Seebeck-Peltier transition in cell membrane?" *Entropy*, vol. 22, no. 9, p. 960, Aug. 2020.

[2] J. C. Eissenberg, "Epigenetics: Modifying the genetic blueprint," *Missouri Med.*, vol. 111, no. 5, p. 428, 2014.

[3] M. Christmann, M. T. Tomicic, W. P. Roos, and B. Kaina, "Mechanisms of human DNA repair: An update," *Toxicology*, vol. 193, nos. 1–2, pp. 3–34, Nov. 2003.

[4] C. F. Guerra, F. M. Bickelhaupt, J. G. Snijders, and E. J. Baerends, "Hydrogen bonding in DNA base pairs: Reconciliation of theory and experiment," *J. Amer. Chem. Soc.*, vol. 122, no. 17, pp. 4117–4128, May 2000.

[5] R. C. Olby, *The Path to the Double Helix: The Discovery of DNA*. Chelmsford, MA, USA: Courier Corporation, 1994.

[6] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter, "Complementary DNA sequencing: Expressed sequence tags and human genome project," *Science*, vol. 252, no. 5013, pp. 1651–1656, Jun. 1991.

[7] J. J. Yunis, "High resolution of human chromosomes," *Science*, vol. 191, no. 4233, pp. 1268–1270, Mar. 1976.

[8] T. Wang and H. Tang, "The physical characteristics of human proteins in different biological functions," *PLoS ONE*, vol. 12, no. 5, May 2017, Art. no. e0176234.

[9] J. Massonneau, C. Lacombe-Burgoyne, and G. Boissonneault, "PH-induced variations in the TK1 gene model," *Mutation Res./Genetic Toxicol. Environ. Mutagenesis*, vol. 849, Jan. 2020, Art. no. 503128.

[10] L. Xu, B. Barker, and Z. Gu, "Dynamic epistasis for different alleles of the same gene," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 26, pp. 10420–10425, Jun. 2012.

[11] J. E. Kammenga, "The background puzzle: How identical mutations in the same gene lead to different disease symptoms," *FEBS J.*, vol. 284, no. 20, pp. 3362–3373, Oct. 2017.

[12] B.-S. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L.-C. Tsui, "Identification of the cystic fibrosis gene: Genetic analysis," *Science*, vol. 245, no. 4922, pp. 1073–1080, Sep. 1989.

[13] Alliance, Genetic, and New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services, *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals*. Morrisville, NC, USA: Lulu, 2009.

[14] E. H. Baugh, H. Ke, A. J. Levine, R. A. Bonneau, and C. S. Chan, "Why are there hotspot mutations in the TP53 gene in human cancers?" *Cell Death Differentiation*, vol. 25, no. 1, pp. 154–160, Jan. 2018.

[15] J.-C. Bourdon, S. Surget, and M. P. Khoury, "Uncovering the role of p53 splice variants in human malignancy: A clinical perspective," *OncoTargets Therapy*, vol. 7, p. 57, Dec. 2013.

[16] P. Neamatollahi, M. Hadi, and M. Naghibzadeh, "Simple and efficient pattern matching algorithms for biological sequences," *IEEE Access*, vol. 8, pp. 23838–23846, 2020.

[17] W. Rhalem, J. El Mhamdi, M. Raji, A. Hammouch, A. Nabil, N. Kharmoum, and H. Ghazal, "An efficient and rapid method for detection of mutations in deoxyribonucleic acid–sequences," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 278–286, 2020, doi: 10.14569/IJACSA.2020.0110438.

[18] L. Loewe, "Genetic mutation," *Nature Educ.*, vol. 1, no. 1, p. 113, 2008.

[19] A. Telenti and J. di Iulio, "Regulatory genome variants in human susceptibility to infection," *Hum. Genet.*, vol. 139, nos. 6–7, pp. 759–768, Jun. 2020.

[20] Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.

[21] N. Mahdieh and B. Rabbani, "An overview of mutation detection methods in genetic disorders," *Iranian J. Pediatrics*, vol. 23, no. 4, pp. 375–388, 2013.

[22] P. A. Hunt and T. J. Hassold, "Human female meiosis: What makes a good egg go bad?" *Trends Genet.*, vol. 24, no. 2, pp. 86–93, Feb. 2008.

[23] U. Eichenlaub-Ritter, "Parental age-related aneuploidy in human germ cells and offspring: A story of past and present," *Environ. Mol. Mutagenesis*, vol. 28, no. 3, pp. 211–236, 1996.

[24] W. Bateson and G. Mendel, *Mendel's Principles of Heredity*. Chelmsford, MA, USA: Courier Corporation, 2013.

[25] R. L. Milne and A. C. Antoniou, "Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers," *Ann. Oncol.*, vol. 22, pp. i11–i17, Jan. 2011.

[26] V. van Heyningen and P. L. Yeyati, "Mechanisms of non-mendelian inheritance in genetic disease," *Hum. Mol. Genet.*, vol. 13, pp. R225–R233, Oct. 2004.

[27] D. S. Cali, G. S. Kalsi, Z. Bingöl, C. Firtina, L. Subramanian, J. S. Kim, R. Ausavarungnirun, M. Alser, J. Gomez-Luna, A. Boroumand, A. Nori, A. Scibisz, S. Subramoney, C. Alkan, S. Ghose, and O. Mutlu, "GenASM: A high-performance, low-power approximate string matching acceleration framework for genome sequence analysis," 2020, *arXiv:2009.07692*.

[28] Fridman, "Memory classification," Int. Fac. Eng., Lodz Univ. Technol., Łódź, Poland, Tech. Rep., 2016, pp. 1–14. [Online]. Available: http://www.ics.p.lodz.pl/%7B~%7Ddpuchala/CompArch/Lecture_6.pdf

[29] D. Groome, "Chapter 6 memory," 2019, pp. 1–17. Accessed: Jun. 2021. [Online]. Available: https://www.polyteknisk.dk/related_materials/9780789736970_Chapter_6.pdf

[30] T. Kyte, "Memory structures," 2010, pp. 121–164. Accessed: Jun. 2021. [Online]. Available: http://docencia.ac.upc.edu/master/MIRI/NCD/docs/03-Memory Structures.pdf

[31] R. Karam, R. Puri, S. Ghosh, and S. Bhunia, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proc. IEEE*, vol. 103, no. 8, pp. 1311–1330, Aug. 2015.

[32] Z. Ullah, "LH-CAM: Logic-based higher performance binary CAM architecture on FPGA," *IEEE Embedded Syst. Lett.*, vol. 9, no. 2, pp. 29–32, Jun. 2017.

[33] Z. Ullah, K. Ilgon, and S. Baeg, "Hybrid partitioned SRAM-based ternary content addressable memory," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 12, pp. 2969–2979, Dec. 2012.

[34] Z. Ullah, M. K. Jaiswal, Y. C. Chan, and R. C. C. Cheung, "FPGA implementation of SRAM-based ternary content addressable memory," in *Proc. IEEE 26th Int. Parallel Distrib. Process. Symp. Workshops PhD Forum*, May 2012, pp. 383–389.

[35] *Introduction to FPGA Acceleration*, Stemmer Imag. GmbH, Puchheim, Germany, 2017, pp. 1–3. [Online]. Available: https://www.commonvisionblox.com/en/technical-tips/introduction-to-fpga-acceleration/

[36] J. Serrano, "Introduction to FPGA design," *CAS-CERN Accel. School, Course Digit. Signal Process.*, pp. 231–247, 2008.

[37] H. Mahmood, Z. Ullah, O. Mujahid, I. Ullah, and A. Hafeez, "Beyond the limits of typical strategies: Resources efficient FPGA-based TCAM," *IEEE Embedded Syst. Lett.*, vol. 11, no. 3, pp. 89–92, Sep. 2019.

[38] J. Xuan, Y. Yu, T. Qing, L. Guo, and L. Shi, "Next-generation sequencing in the clinic: Promises and challenges," *Cancer Lett.*, vol. 340, no. 2, pp. 284–295, 2013.

[39] B. Schmidt and A. Hildebrandt, "Next-generation sequencing: Big data meets high performance computing," *Drug Discovery Today*, vol. 22, no. 4, pp. 712–717, 2017.

[40] S. Samaddar, R. Sinha, and R. K. De, "A model for distributed processing and analyses of NGS data under map-reduce paradigm," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 827–840, May 2019.

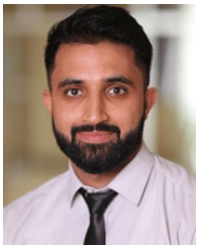
[41] A. Mogi and H. Kuwano, "TP53 mutations in nonsmall cell lung cancer," *J. Biomed. Biotechnol.*, vol. 2011, pp. 1–9, Oct. 2011.

[42] A.-L. Børresen-Dale, "TP53 and breast cancer," *Hum. Mutation*, vol. 21, no. 3, pp. 292–300, 2003.

[43] B. Iacopetta, "TP53 mutation in colorectal cancer," *Hum. Mutation*, vol. 21, no. 3, pp. 271–276, Mar. 2003.

[44] D. P. Guimaraes and P. Hainaut, "TP53: A key gene in human cancer," *Biochimie*, vol. 84, no. 1, pp. 83–93, Jan. 2002.

[45] *IARC TP53 Database*. Accessed: Jun. 2021. [Online]. Available: https://p53.iarc.fr/p53Sequence.aspx



MUHAMMAD IBRAHIM received the B.Sc. degree in electrical engineering and the M.Sc. degree in electrical communication from the CECOS University of IT and Emerging Sciences, Peshawar, Pakistan, in 2017 and 2020, respectively. In 2018, he started working as a Lab Engineer with the Department of Electrical Engineering, CECOS University of IT and Emerging Sciences. He has published two peer-reviewed international IEEE conference papers. His research interests include embedded systems, FPGA-based systems, biomedical engineering, and artificial intelligence.

OMER MUJAHID received the B.Sc. degree in electrical engineering and the M.Sc. degree in electrical communication engineering from the CECOS University of IT and Emerging Sciences, Pakistan, in 2013 and 2017, respectively. He is currently pursuing the Ph.D. degree with the University of Girona, Spain. He is also associated with the Model, Identification and Control Engineering (MICE) Lab, Institute of Informatics and Applications (IIiA), University of Girona. His research interests include machine learning for diabetes decision support, deep learning for biomedical data simulation, and reconfigurable computing.

NAJIB UR REHMAN received the B.Sc. degree in electrical engineering and the M.Sc. degree in electrical engineering (communication) from the CECOS University of IT and Emerging Sciences, Peshawar, in 2013 and 2019, respectively. He joined the Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, as a Lab Engineer. He has published two IEEE conference papers. His research interests include embedded systems, FPGA-based systems, and artificial intelligence.

AZHAR QAZI received the B.Sc. (Hons.) and M.S. degrees in electrical engineering (communication) from the University of Engineering and Technology, Peshawar, Peshawar, Pakistan, in 2006 and 2014, respectively, and the Ph.D. degree from the Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar, in 2020. He is currently with the Department of Electrical Engineering, CECOS University of IT and Emerging Sciences. His research interests include designing fast updating and mapping algorithms for SRAM-based CAMs on FPGA, memory management in fast lookup algorithms, and the development of intelligent algorithms for traffic flow.

ZAHID ULLAH has served as an Associate Professor and the Chairperson for the Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar, Pakistan. He got more than 12 years of experience in teaching and research while working in different places, which include, in addition to above, the City University of Hong Kong (CityU Architecture Lab for Arithmetic and Security), Hanyang University, South Korea (Reliable and High-Speed Computing Laboratory), the FAST National University of Computer and Emerging Sciences, Peshawar, the Peshawar College of Engineering, Peshawar, Pakistan, the University of Engineering & Technology Mardan, Mardan, Pakistan, and Siemens (Pakistan) Engineering Company Ltd. He is currently serving as an Assistant Professor and the Head for the Department of Electrical and Computer Engineering, Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Haripur, Pakistan. He is an Expert in outcome-based education system (OBE) of the Washington Accord. Further, he has also completed four months training on ''Fachhochschule Teaching and Management'' in two Austrian Universities–FH Joanneum, Graz, and MCI, Innsbruck. He has supervised two Ph.D. students and more than 20 M.S. students. He has around 50 journal and conference papers and has got six patents (U.S. and Korean) on his name in the field of FPGA-based TCAMs. His research interests include low power/high speed CAM design on FPGA, re-configurable computing, pattern recognition, embedded systems, and image processing. He was awarded scholarship for M.S. studies at Hanyang University, South Korea, by the Higher Education Commission, Pakistan, and for Ph.D. studies by the City University of Hong Kong. He also received two times Best Researcher Award (CECOS University), two times Outstanding Academic Performance Award, two times Research Tuition Fee Scholarship, and two times Conference Grant (City University Hong Kong).

TAMA FOUZDER received the bachelor's degree from the Discipline of Electronics and Communication Engineering, Khulna University, Bangladesh, in 2006, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, in 2016. She worked as a Research Assistant with the EPA Centre, Department of Electronic Engineering, City University of Hong Kong. She has around nine years teaching experience as a Lecturer with the Department of Electronic and Telecommunication Engineering, University of Development Alternative, Bangladesh. She is currently working as an Assistant Professor with the Department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh, Bangladesh. She has published ten international journal articles and ten peer-reviewed international conference papers. Her research interests include electronic product reliability and electronic packaging, lead-free nano composite solders and final surface finish, FPGA based design, antenna design, and biomedical engineering. She served as a reviewer for some international journals and international conference papers.