International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 4, July-Aug-2020, ISSN (Online): 2395-566X

Genome Sequencing and Analysis of Cancer Codons

Shailesh.D, Anooja Ali, Vishwanath R Hulipalled, Harshitha KB, Swetha Sivakumar, Manjunath C
School of Computing and Information Technology
REVA University Bengaluru, India
shaileshshettyd@gmail.com, anoojaali@gmail.com, vishwa.gld@gmail.com, sanjanakudo0405@gmail.com, swetshiv@gmail.com, manju2403897@gmail.com

***Abstract* – Genome sequencing helps to identify the variation in genomic structure or to detect the new genomic sequences over the population. This sequence alignment detects the preserved interactome unit among organism. It evaluates the phylogenetic distance among organism of same species or different species and thereby detecting the functional domains, polymorphisms, and mutations. Sequence alignment is the arrangement of the sequences of deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or macromolecule to spot regions of similarity which will be a consequence of practical, structural or organic process relationships between the sequences. In this paper, we detect mutation and the presence of cancer codon using pattern matching and sequence alignment. The sequences are preprocessed using Boyer Moore and k-mer indexing algorithms. Mutations present in the sequence is detected with local and global alignment with Naive approximation. Gene sequencing helps to understand and evaluate the genomic characteristics of an organism at a lesser cost and with great coverage.**

***Keywords*– Boyer-Moore, Genome Sequencing, Mutation, Phylogeny, Polymorphism, Sequence Alignment.**

I. INTRODUCTION

The genome sequencing is an emerging technique applicable in the field of bioinformatics. It is also been used in many other fields such as anthropology, biotechnology, forensic science, etc. Human genome mapping and sequencing will give several potential information about human diseases. Identifying the frequently occurring complex genomic pattern is the primary solution for several disease identification. The existing social and ethical challenges in genome mapping are addresses by several scientists [1].

Biological sequences are functionally analyzed by multiple sequence alignment. It gives an insight to the sequence family, sequence structure and their relationships. During every multiple sequence alignment, a given sequence is compared to various set of sequences from the connected organisms. Homology of the compared sequences reveals their evolutionary relationships. The amino acid sequence alignment and its analysis is the center to most of the biological applications. Reliable and efficient algorithms must be used for alignment. Local and global alignment based on dynamic programming are the popular methods of alignment [2].

Cancer genomic study has revealed the high intra as well as genetic heterogeneity. The identification of cancer driver genes and its mutations is a major central problem in the cancer research. When the DNA that monitors the cellular function changes, the cell divide and grows, thereby acting as carcinogens. The Cancer Genome Anatomy Project (CGAP) launched in 1997, imprinted the RNA sequences present in tumor cells [3]. Cancer genome sequencing facilitates the oncologists to detect the specific variations underwent by a patient in the development of cancer [4].

In this paper, we perform DNA sequencing to detect cancer codons and mutation. The paper is arranges as the next section is the literature survey. Following this system architecture is presented and then the results and discussion. Using these results, we have the potential to discover the contributions of genomic variants to various organisms' health and diseases.

II. LITERATURE SURVEY

DNA sequence is the representation of genetic code contained with an organism. Genomics research often requires gathering data about genomic variation, phenotypes, demographics, and exposures. String searching algorithms finds the patterns that are common in a sequence. The common string matching algorithms are Naïve string search algorithm, Rabin- Karp algorithm, Knuth–Morris–Pratt algorithm, k-mer indexing, and Boyer–Moore string search algorithm [5]. The performance of the algorithms are measured using time taken for single or multiple word search, number of iterations and accuracy.

Naïve string search algorithm is a brute force method that doesn't require any string preprocessing method [6]. As it consider only one position at a time, it is inefficient. To search for a pattern p with m as mutual length for a text of length n, Boyer-Moore Algorithm is the most efficient

Algorithm for general matching and its primary function is to gain more information by matching the pattern from right to left, which makes faster run-time [8]. K-mer indexing produces k-mers based on the pattern required and are typically used during sequence assembly, but can also be used in the sequence alignment. K-mer indexing can also be used as a first stage analysis before alignment. It is used to find DNA barcoding of species, de novo sequence assembly, detect genome miss- assembly, and estimate the genome size [9].

Pattern Matching algorithms are mainly two types, Approximate and Exact pattern matching. Approximate pattern matching is primarily useful in finding approximate occurrences of Pattern P in a text T using edit distance matrix. Traceback of the Pattern Sequences has vertical backtrace which indicates the missing character and diagonal value increment leads to the mismatch. Distance-based hamming method can accept characters mismatches in the arrangement, gives different performance results depending upon several compared patterns. Exact pattern matching is useful in finding exact matches of a pattern P in the text T at offsets[10].Let x=|P| and y=|T|, the greatest number of character comparisons possible in the algorithm is x(y-x+1) and the least character comparisons is y-x+1.The Exact-pattern matching or Naive pattern matching is the most efficient matching algorithm in terms of computational efficiency.

DNA sequencing is the process of determining the exact order of nucleotides within a DNA molecule. DNA sequencing alignment is a representation of similarity between two or more sections of the genetic code. There are mainly two types of sequence alignment namely, global and local sequence alignment. Needleman-Wunsch algorithm is the common global alignment technique which is primarily based on dynamic programming, to find alignment between two sequences which are similar in length as well as similar across the width [11]. It employs dynamic programming to determine optimal local arrangements for sequence similarity between the pattern sequence and text sequence. This type of sequence alignment helps in finding the most similar pair of substrings from Pattern P and Text T. The Scoring matrix will have the matches as Positive 1 where mismatches and gaps have penalty value 0.

The mutational differences between two different gene sequences to determine the mutated cancer gene. We compare the efficiency of each algorithm in each module to ensure accuracy in finding mutations. We increase the efficiency of finding mutations with primary techniques namely, indexing, pattern matching, sequence alignments and comparison of FASTA files to find out variation, leading to the increase in accuracy of finding cancer genome mutation.

III. SYSTEM ARCHITECTURE

Sequence Preprocessing- Preprocessing is done with K-mer Indexing Algorithm and Boyer-Moore Algorithm. Pattern Matching- Naive Pattern Matching is used to check if pattern string matches with the text string whereas approximate string matching finds the strings that match the pattern using the distance(mismatches).

Sequence Alignment- We focus mainly on pairwise sequence alignments where the runtime efficiency is calculated by comparing two sequences Pattern P and Text T. Both the mutations will be depicted as gaps and aligned sequences as matches. Mutation Detection-The mutations of the samples are found by parsing the normal and treated samples. Zip command is used to compare the mismatch and its sum function is used to find the mutations of the samples.

1. Sequence Preprocessing

Sequence preprocessing is performed using k-mer. The first step in any k-mer analysis is the generation of a profile, which is constructed by the indexing algorithm. K-mer refers to all the potential substrings of a string of length k [13]. The efficiency of the algorithm is enhanced by encoding the DNA string in binary. Following, the binary encoded k-mers are used as the index of a count table. This can be achieved by the concatenation of the binary code for each nucleotide in a given DNA string. This procedure eliminates the need to store the actual k-mer sequences since they can be retrieved from decoding the offset in the count table. The binary code for each nucleotide is chosen in such a way that the complement of the nucleotide can be intended using the binary NOT operator. The indexing algorithm returns a profile that holds observed counts for all possible substrings of length k that can be stored for other analyses [14].

Boyer-Moore Algorithm focuses mainly on two processes that is Alignment which applies from left to right whereas string search applies from right to left and sequences are skipped until match is found among sequences P and T [15]. K-mer assigns the indexes to each nucleotide and depicts the range of matched string whereas Boyer-Moore depicts the exact positions of matched nucleotides during align and string search operations.

2. Sequence Alignment

In the Smith-Waterman (local), alignment traceback is performed from the highest score element in the matrix (the highest scoring end pairs).The edit distance penalized all the different kinds of edits that we might find at same amount rate whether there is a mismatch or insertion or deletion. The edit distance would penalize unit of one and substitutions occur between sequence alignments [16]. Fig. 2 and 3 shows the local and global alignment pattern.

International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 4, July-Aug-2020, ISSN (Online): 2395-566X

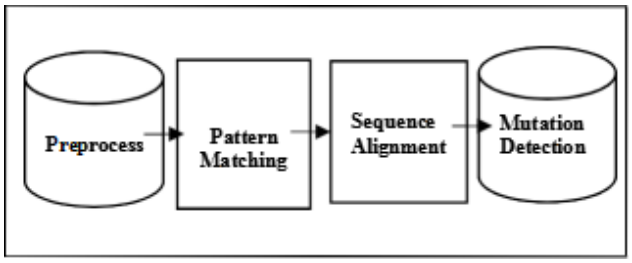

Fig. 1. System Architecture.

```
          tccCAGTTATGTCAGgggacacgagcatgcagagac
                    ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Fig. 2. Local Alignment.

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

Fig. 3.Global Alignment.

Global Sequence alignments are useful in finding the similarity nucleotides among pattern sequences P and text sequences T and the gaps define the mutations between them. The procedures followed in the Smith-Waterman(local) algorithm is as follows(1) Assign the values of initial gaps sequence as binary 0.(2) Later, Consider the edit Distance among vertical, horizontal and diagonal blocks where the values are assigned already.(3)Calculate the maximum values among the distance penalty score and continue filling the penalty scores till the last nucleotide in a text T.(4)Traceback method starts from last nucleotide to the first nucleotide in text T where gaps in traceback are identified as mutated nucleotide in Pattern P.

In Needleman-Wunsch (global) algorithm, the match, mismatches, and gaps possess integer specific values in penalty matrix. The matches are assigned with positive 1(+1), mismatches as negative 1(-1) and gaps are assigned as negative 2(-2). The Score filling approach and Traceback follows same approach as local sequence alignment algorithm. The efficiency of the Global sequence alignment is exceptionally higher in terms of runtimes which is calculated and then call global alignment along with pattern and text sequence.The alignment depicts not only the matched nucleotide sequences but focus majorly on the mismatches or mutated sequences which helps in finding mutations in alignment representation.

The work is carried out with Bio python. Bio python has capacities for analysis of gene sequences, detecting motifs present in sequences, phylogenetic and sequence alignment. We use Fluent DNA for Visualizing the DNA sample of a normal and mutated sample. Adenine(A), Cytosine(C), Guanine(G), Thymine(T) is assigned with colors to differentiate and N represents the mutation of the sample.

IV. RESULTS AND DISCUSSION

Genome Analysis of any Organism can be detected with the help of the program at a very fast rate and under high accuracy. Information provided by the genomic sequencing can be a value to the future generation for mutation detection.

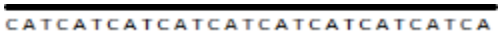

Fig .4. DNA code.

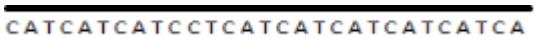

Fig .5.Mutation detected.

Fig 4 denotes the amino acid sequence. When any deletion or insertion or change occurs for of any nucleotide in the sequence, it is concluded about the presence of mutation. The DNA bases CAT produces the amino acid Histidine, the crucial amino acid that produces histamine. When sequences changes from CAT to CCT, the corresponding amino acid is Leucine. Protein biosynthesis is performed by Leucine. Any variation in amino acid produces amino acid variation. This is denoted in fig 5.

The genome sequencing compares the genetic DNA with the normal DNA to detect the anomalies in the sequencing pattern and deduces the mutation if found through cascade testing. Cascade testing with other family members provides crucial evidence required for Genome Sequencing. The DNA can be analyzed using FASTA Visualization format helping to produce accurate results.

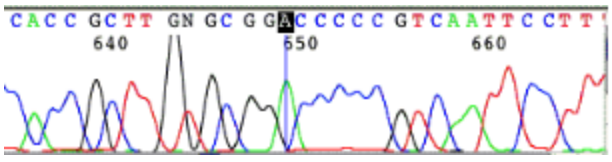

Fig. 6: mutation detected.

Fig.7.Mutation detected for cervical cancer sequence AAK97314.1

>AAK97314.1 cervical cancer proto-oncogene 2 [Homo sapiens]
MQAVRNAGSRFLRSWTWPQTAGVVARTPAGTICT GA
RQLQDAAAKQKVEQNAAPSHTKFSIYPPIPGEESSL RA
GKKFEEIPIAHIKASHNNTQIQVVSASNEPLAFASCG TE
GFRNAKKGTGIAAQTAGIAAAARAKQKGVIHIRVV V
KGLGPGRLSAMHGLIMGGLEVISITDNTPIPHNGCR PR KARKL

Fig.8. Cervical cancerous protein sequence.

DNA trace file known as ABI is used for mutation detection. It compares the DNA sequences and detects mutation. Fig. 6 shows that mutation is detected at bases 644 and 650.The spikes obtained at the base position indicates this. Fig.7 indicates the mutation detected when an infected cervical cancer sequence was sequence was tested with healthy sequence. The four colors indicates the four bases (Adenine, Thymine, Guanine, and Cytosine). Any other color indicate mutation. The protein sequence used for detection is given in fig. 8. AAK97314.1 is the accession number of the sequence [18]. The sequence is obtained from National Centre for Biotechnology Information (NCBI). The vital proteins E1 and E2 are main causes for infection in cervical cancer. Consensus motif, AACNAT is present in E1 helicase which indicates the infection and thereby diagnosing as cervical cancer.

V. CONCLUSION

Genome sequencing helps in extracting useful part of sequences which are responsible for cancer mutations from DNA. Genome Analysis involves the step by step procedures such as parsing, manipulations, complement, transcription and translation, alignments and mutation detection using efficient algorithms. In our paper, Genome sequencing and analysis is performed using initial sequence preprocessing and later the pattern matching, sequence alignment and mutation detection for each cancer genome of various organisms. These series of steps from indexing to alignment effectively detects mutations present in any gene sequence. Scientific information regarding the Genome Sequencing can be obtained with the potential medical implications that can be used to detect the anomalies occurring in the cells of organisms.

REFERENCES

[1]. Huang, Bevan E., Widya Mulyasasmita, and Gunaretnam Rajagopal. "The path from big data to precision medicine." Expert Review of Precision Medicine and Drug Development1.2 (2016): 129-143.
[2]. Didelot, Xavier, et al. "Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks." Molecular biology and evolution 34.4 (2017): 997-1007.
[3]. Strausberg, Robert L., et al. "The cancer genome anatomy project: building an annotated gene index." Trends in Genetics16.3 (2000): 103-106.
[4]. Massagué, Joan, and Anna C. Obenauf. "Metastatic colonization by circulating tumor cells." Nature 529.7586 (2016): 298.
[5]. Fournier-Viger, Philippe, et al. "A survey of sequential pattern mining." Data Science and Pattern Recognition 1.1 (2017): 54-77.
[6]. Chhabra, Tamanna, and Jorma Tarhio. "Order-preserving matching with filtration." International Symposium on Experimental Algorithms. Springer, Cham, 2014.
[7]. Lubis, Andre Hasudungan, Ali Ikhwan, and Phak Len Eh Kan. "Combination of Levenshtein distance and Rabin-Karp to improve the accuracy of document equivalence level." International Journal of Engineering & Technology 7.2.27 (2018): 17-21.
[8]. Faro, Simone, and M. Oğuzhan Külekci. "Fast and flexible packed string matching." Journal of Discrete Algorithms 28 (2014): 61-72.
[9]. Quinn, Jeffrey J., and Howard Y. Chang. "Unique features of long non- coding RNA biogenesis and function." Nature Reviews Genetics 17.1 (2016): 47.
[10]. Ballena, Kaliuday, D. Satyanvesh, and P. K. Baruah. "GenSeeK: A Novel Parallel Multiple Pattern Recognition Algorithm for DNA Sequences." Intelligent Computing, Networking, and Informatics. Springer, New Delhi, 2014. 1001-1006.
[11]. Bray, Nick, Inna Dubchak, and Lior Pachter. "AVID: A global alignment program." Genome research 13.1 (2003): 97-102.
[12]. Aluru, Srinivas, and Nagakishore Jammula. "A review of hardware acceleration for computational genomics." IEEE Design & Test 31.1 (2014): 19-30.
[13]. Restrepo, Juan Manuel Ciro, Andrés Felipe Zapata Palacio, and Mauricio Toro. "Assembling sequences of DNA using an online algorithm based on de Bruijn graphs." arXiv preprint arXiv:1705.05105 (2017).
[14]. Wu, Yu-Wei, et al. "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation- maximization algorithm." Microbiome 2.1 (2014): 26.
[15]. Chhabra, Tamanna, and Jorma Tarhio. "A filtration method for order- preserving matching." Information Processing Letters116.2 (2016): 71-74.
[16]. Gawad, Charles, Winston Koh, and Stephen R. Quake. "Single-cell genome sequencing: current state of the science." Nature Reviews Genetics 17.3 (2016): 175.
[17]. Luu, Vinh-Trung, et al. "Using glocal event alignment for comparing sequences of significantly

different lengths." International Conference on Machine Learning and Data Mining in Pattern Recognition. Springer, Cham, 2016.

[18]. Van Doorslaer, Koenraad, and Robert D Burk. “Evolution of human papillomavirus carcinogenicity.” Advances in virus research vol. 77 (2010)