Motif Detection in Cellular Tumor p53 Antigen Protein Sequences by using Bioinformatics Big Data Analytical Techniques

Tariq Ali[1], Sana Yasin[2], Umar Draz[3], Tayyaba Tariq[5], Sarah Javaid[6]
CS. Department
(CIIT) Sahiwal, Pakistan

M. Ayaz Arshad[4]
SCNC Research Centre
UoT, Tabuk,
Kingdom of Saudi Arabia

***Abstract*—Due to the rapid growth of data in the field of big data and bioinformatics, the analysis and management of the data is a very difficult task for the scientist and the researchers. Data exists in many formats like in the form of groups and clusters. The data that exist in the group form and have some repetition patterns called Motifs. A lot of tools and techniques are available in the literature to detect the motifs in different fields like neural networks, antigen/antibody protein, metabolic pathways, DNA/RNA sequences and Protein-Protein Interactions (PPI). In this paper, motif detection is done in tumor antigen protein, namely, cellular tumor antigen p53 (Guardian of the protein and genome) that regulate the cell cycle and suppress the tumor growth in the human body. As tumor is a death causing disease and creates a lot of other diseases in human beings like brain stroke, brain hemorrhage, etc. So there needs to investigate the relation of the tumor protein that prevents the human from not only brain tumor but also from a lot of other diseases that is created from it. To find out the gap between the motifs in the tumor antigen the GLAM2 is used that detects the distance between the motifs very efficiently. Same tumor antigen protein is evaluated at different tools like MEME, TOMTOM, Motif Finder and DREME to analyze the results critically. As tumor protein exists in multiple species, so comparison of homo tumor antigen protein is also done in different species to check the diversity level of this protein. Our purposed approach gives better results and less computational time than other approaches for different types of user characteristics.**

***Keywords*—Bio-informatics; motif detection; guardian protein Tp53; DNA; tumor antigen; cancer; un-gapped motifs; MEME**

I. Introduction

Big data became an active research from the last few years due to its immeasurable range of applications. Due to rapidly increasing trends and interest of research in this domain, there are many improvements have been done in this field that become famous among the research society due to manage the large amount of data that cannot be handled by the traditional databases [1]. Instead of the momentous work on the motif discovery, motif detection in tumor proteins remains a difficult task for computer scientists and biologists. Lot of encouraging tools and algorithm is purposed in this field to make progress. Huge attempts have been done for the enlargement of the computational techniques for the identification of the sequence motifs in proteins. In the field of bioinformatics motif detection is an exigent problem due to the variety of protein motifs. In [2] author divides the biological motifs in three classes. Each class contains different type of motifs like the class ‘A’ contains the motifs that are in small size and appear at the functional sites of the biopolymer, cleavage and binding sites is the example of such type of motifs. The class ‘B’ contains large size motifs that are frequently crop up due to the divergent evolution and these motifs are highly associated with spherical structural domain. The recurring motifs fall in the class ‘C’ and these motifs are appearing due to the innovative recent replications. Due to diverts and complex nature of each class it’s too much difficult to tackle all type of motif through single motif searching method (SMSM). There are multiple techniques to discover the over-represented motifs in the protein sequence to maximize the expectation in the sequence, but these techniques do not give the appropriate result. In this paper, graphical approach is used to detect the motifs in the tumor protein p53 that is the ***“guardian of the proteins and genome”*** because of its role in conserving stability by preventing genome mutation [3]. Among the field of protein-protein interactions the protein p53 has immune effect in the medical health sciences such as controls the oxidative stress, DNA damage; manage the functionality of ribosomal dysfunction and Hypoxia. In order to determine the rate of some metabolic and anabolic reactions when a cell protects from one step to the path for cancer and different diseases, therefore it is also called the tumor protein Tp53. The unique features and functionalities of the tumor protein p53 are shown in Fig. 1.

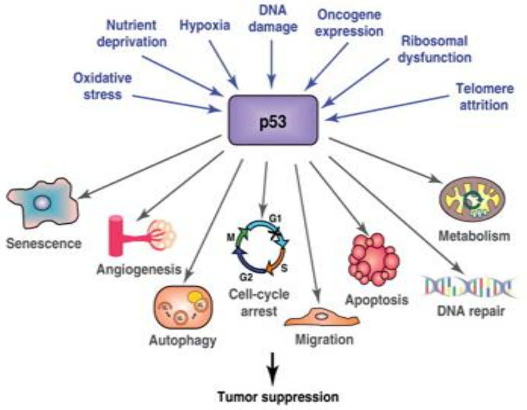


Fig. 1. Characteristics of Tumor protein p53.

The methods that were used in past for the motif detection was very slow and they search the motif of equal length. This motif does not work accurately in the class identification process in which motifs has different length and type. Motifs are located at innumerable distances in the protein sequence so to discover the gapped and un-gapped motifs is an essential task in the field of bioinformatics and big data. The importance of the gapped motifs is demonstrated also by the fact that many databases exist that contains motifs like PROSITE and ELM that contain gapped and different length motifs [4]. Newly purposed graph-based motifs detection technique efficiently searches the gapped and un-gapped motifs in the protein sequence. The tumor protein p53 data set that is selected for the motifs detection performs a momentous role in the body to control the cell cycle and apoptosis. The motifs that exist in the tumor protein p53 are minor persistent patterns that are accredited to have a conventional task. Defective p53 could be conceivably allowing the abnormal cells to promulgate that resulting in a tumor. As well as 55% of all human tumors comprise p53 mutants but in common cells of the human body, p53 protein level is small [5]. There are a lot of factors in the homo species that increase the p53 protein ratio like stress signal and DNA damage. There are multiple functions of p53: growth arrest, DNA repair and apoptosis (cell death) [6]. Tumor suppression becomes reduced in the human body if the p53 protein becomes damaged. A disease Li-Fraumeni syndrome occurs in the childhood if people inherit only one functional copy of the p53. The proposed research addresses the following issues:

- Motif detection in tumor protein
- Investigate the suitable parameters for the detection of motifs
- Reduce the tumor ratio in homo species by analyzing the proteins tumor that suppresses the tumor
- Identification of a suitable and match motif in the tumor protein
- Comparison of tumor protein motifs in varied species

In Fig. 2, the structure of different tumor proteins is represented with the root and leaf motifs hierarchy. The alignment of different motifs of tumor protein at some different level is shown as a red label. This level further divides the motif roots where all the propagation of the motifs is present. In another way, the motif is basically the sub-part of motif root.

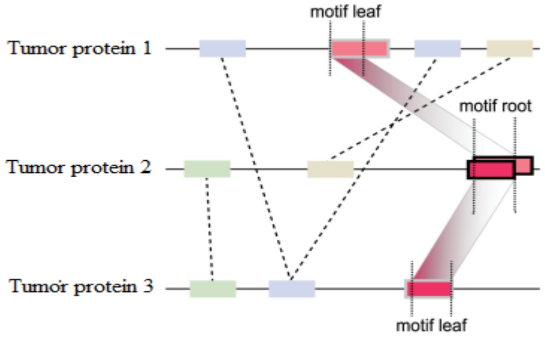


Fig. 2. Alignment of Motifs in the Tumor protein p53.

The major objective of this paper is to analyze the detection of tumor protein in the form of Motif Detection Algorithm (MDA) with the help of some suitable parameters. Furthermore, this research gives an idea of how it is possible to reduce the ratio of tumor proteins that suppress the tumor. After this, with the help of different tools, the two categories are assigning like residue motifs and site motifs that help to identify the matched motif in a different location. At the end, for the reliability of the results and its efficiency, the comparison is done between different species and identifies the number of matched motifs. The same protein dataset of p53 has been evaluated in different tools and Motif Finder technique, for the detection and evaluation of gapped motifs the GLAME2 is used and for the detection of un-gapped motifs, the MEME data bank is used. To perform the comparison analysis between the Tp53 the Motif Enrichment Analysis (MEA) is used that determines the position of best sites of the motifs against its possible probability of the detective motifs among the given data sets of the protein. Finally, after the comparison, the motif between different species the residue and site-based motifs are categorized.

Rest of this paper is organized as: section II is discussed the related work. Section III discussed the motif representation. Section IV describes the purposed methodology. Section V contains Motif Detection Algorithm. Section VI deal with the results and discussion. At the end, the conclusion is represented in Section VII.

II. Related Work

A lot of research has been done at the motif detection in the field of big data and bioinformatics, but still more attention is required in this field. Recently, the research on efficient mining of previously unfamiliar, recurrently emerging patterns has received much attention in the field of medical health sciences [7]. With the advancement of technology and trend of social media; the amount of data is growing very rapidly. This data is spread across different places, in different formats, in large volumes ranging from Gigabytes to Terabytes, Petabytes, and even more. Today, the data is not only generated by humans, but enormous amounts of data are being generated by machines and it surpasses human-generated data [8]. This size aspect of data is referred to as volume. In the Big Data world lot of work has been done but there still more work is pending to the amount of different data aspects like in Bioinformatics Big Data (BBD). Motif detection is some of the highly focused topics among the researchers in the big data research community. These motifs are useful for various time-series and data mining tasks. The relation between DNA and protein is a key motivating force. Binding of protein-sites and the specially targeted proteins are two important moves to understand the concept of biological activities [9]. A lot of techniques that gives high-throughput have recently purposed that try to enumerate the similarity between proteins motifs and protein [10]. In spite of the strong achievement, these techniques have some limitations and go down towards the strict classification of motifs. As a result, need further critical analysis of protein and protein sequences to dig out useful and modifiable information from a stack of strident and raw data. In [11] Motif Mark Algorithm (MMA) is purposed to find out the regular motifs in the protein sequence. This algorithm based on the graph theory and machine learning that finds binding sites of protein sequences. It also analyzes investigational data that is derived from universal protein against two of the most precise motif detection methods [12]. Gene mapping has been considered as one of the challenging tasks for researchers who belong to the field of bioinformatics and data analysis. Previously such tools are developed which are used for gene analysis and gene mapping for example MAPMAKER [10]. MAPMAKER has been applied to the construction of linkage maps in a number of organisms; including the human beings [13]. Mutations [14] in its structures can cause various diseases for this purpose simulation in proteins can reveal many new structures. One of the other techniques introduced for protein unfolding is Steering MD [14]. In [15] analysis is done on the whole genomes to find out the repetitive protein sequences called non-B motifs. These motifs are capable to predict the non-canonical structure of the protein and can autonomously report for deviation in mutation density. Graph visual motifs are also helpful for distinguish between applications protocol and to determine the known behavior of unlabeled traffic [16]. The most widely used tool for motif discovery is the MEME. MEME is a complete suite and performs a series of operations on the dataset thus discovering, analysis, finding enrichment and comparison with the existing motif databases [17]. Some other tools like DMINDA; Ensemble Genome Browser also performs a sequence of operations [18]. The combined effort of p53 and p63 in some Differential composition of DNA-binding sites may contribute to distinct functions of these protein homologs in some different species. To identify the legends and nuclei of the p54 proteins the SELEX (systematic evolution of legends by exponential enrichment) tool has been used [19]. To arrange the sequence in some protein input the long chain of the sequence has been used, for example; AGTGCGGCCGCTCAGGTTGACTTCCCCGCG.

In Western Bolt Analysis (WBA) [20]-[22] take the data sets of p53 and p63 for the proper cure of cancer in Health-Nutrient laboratory, they found that the p53 is most effective and the dominant parameters that play part and parcel role in the disease of Cancer. So, its need to investigate the p53 tumor protein at different tools to find out the exact relationship with the different disease that plays our role in the sequence. In our proposed algorithm the p53 Motif Detection Algorithm (P53MDA) detects the gapped and un-gapped motifs.

III. Motif Representation

Graphical based approach to find the gapped and un-gapped motifs in the sequence of the protein is in the form of regular expression is presented as:

$$R1 - p\ (n1;\ m1) - R2 - p\ (n2;\ m2) - \ldots\ldots\ldots\ldots - Rr$$

R1 = Base Class
N1 = Least number of the base class
M1 = Most number of the base class

Where ‘R1’ is an un-interrupted sequence with $1 \leq i \leq r$ of amino acids that are called components, while *-p (n1; m1)-* represents a gap of length at least number ‘*n1*’ and at most ‘*1*’. There are three major types of motifs depending on the size of the gap, the first one is contiguous motifs, these motifs have no gaps between them and ‘*n1*’ and ‘*1*’ values are zero ‘*n1*’ = ‘*m1*’ =0 (for un-gapped motifs). The motifs that contain gaps called rigid motifs that fall into the second category of the motifs. The length of the rigid motifs is fixed i.e., ‘*n1= m1*’ for all $1 \leq i \leq r - 1$. The third category of motifs is flexible gap motifs, these motifs contains different size gaps between the two motifs, i.e., *‘n1 ≤ m1; for all 1 ≤ i ≤ r - 1.*

IV. Methodology

The p53 tumor suppressor is implicated in cell cycle control, DNA repair, explicative sequence and programmed cell death. In-activation of the p53 contributes to the wide range of human tumors; including Glial neoplasm’s. Due to its lot benefits and features its need to analyze the tumor protein more critically. In this paper, the proposed algorithm is introduced to detect the motifs of different lengths with gaps and without gaps through motif detection algorithm. The sequence analysis of the tumor protein p53 is performed by using MEME tool. To find the rigid motifs in the tumor protein sequence the GLAME2 tool is used that is very efficient for the detection of gapped motifs in the sequence. Tumor protein Sequence clusters are downloaded from UniProt database. The sequences of p53 are taken as a class of homo species. These sequences have their own significance in genomics. They were selected due to the unique feature of being the *“guardian of the genome”* and example of mutation caused in Homo sapiens. P53MD algorithm is used for finding the motif within the sequences. Discovered Motifs are shortest motifs that found out, than compared using TOMTOM and a resultant table is derived according to the number of matches along with the PROTEINS motifs are found using DMINDA tool box. The motif which is found in maximum number of sequences is compared with the existing database and results are derived. Simulation parameters for the proposed work are discussed in Table I. All the simulation is done with the help of proposed p53MDA and the alignment of this work is considering for both random and discreet. The reason is that for difference between the residue and site motif it is necessary to take the data set is evaluated for some random and discriminative fashion.

Up to our best knowledge this work is firstly done on the basis of both data types format. To select the discovery mode discriminative then the number of protein are aligned is some order, otherwise random order is apply. The novelty of this work is not only the detection of motif inside the protein but also provides the detail comparison among different available tools.

TABLE. I. GENERAL SIMULATION PARAMETERS OF P53 MDA AT MEME, DMINDA, DREME, TOMTOM & MOTIF FINDER

Simulation Parameters	Value
Discovery Mode	Random/Discriminative
Standard Custom Alphabets	Protein Sequence
Expected Motif Distributed	Zero/One Occurrence per sequence
No. of Motifs/Alignment	1000-5000
Strands Identification	ON
Presence of strands	+
Absence of strands	-
Length of Protein	5,000 for Each Protein/Sequence
Computational Timing	Depends on Input Sequence e.g 10 and 15 Seconds Each
Start/End Value	ON
Motif Score/Enrichment	ON
Best Motif Value	Mentioned
Location/Position of Motif	ON
X dimension of topography	1000
Y dimension of topography	4500

V. P53MDA FOR GAPPED AND UN-GAPPED MOTIFS

A. *Pseudo code*

Algorithm 1: p53 Motif Detection Algorithm

In this algorithm, Tumor protein sequence is taken as an input and novel gapped and un-gapped motif are detected that have fixed and variable length.

Input: 'N' numbers of Tumor protein sequence of TP53
Output: Gapped and Un-gapped motifs with variable length

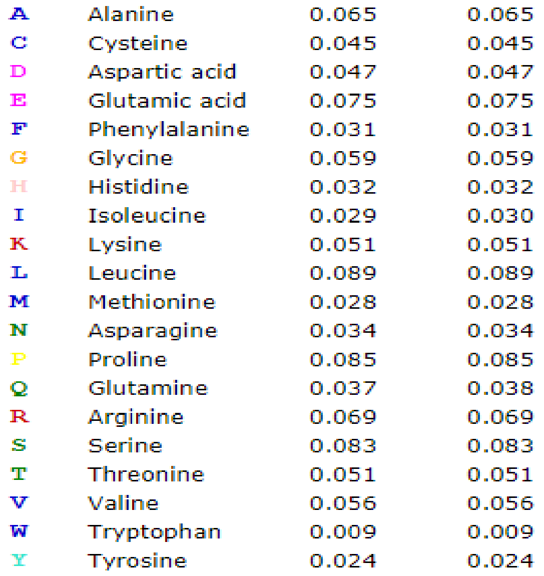
```
Input tumor protein sequence
For s=1,…………,S-1 do
    For i=1,…………,I-s do while {
        For j=1,…………,J-s-i do while {
N1 = 0 // initialize the gap of length at least domain of the motifs variable (selected) from zero
M1 != 0 // initialize the gap of length at most number variable (unselected) from zero
for (s in 1: mid) // this loop runs from 1 to mid for selecting motifs from the first segment of input sequence
{
for (i in 1:2)   // inner for loop is running from 1 to 2 times used for locations
{
for (j in 1:1/2:2)   // inner for loop is running from 1 to half and half to 2 times used for locations
{
k=1
if (s [i, j] = =0 OR a [i, j=j+1] = =0)
{
gapped=gapped+1 // counts the gapped motifs from segment 1
gapped
}
else
{
Un-gapped =un-gapped+1 // counts the un-gapped motifs from segment 1

end
    }
  }
}
```

Fig. 3. High-level pseudo code for P53MDA.

VI. RESULT AND DISCUSSION

In this section, the graphical results have been displayed against the tumor proteins by using different tools with different parameters. Every homo species has nucleus, DNA, RNA, genes and lot of chromosomes. The genes that want to express motifs create their replication in the chromosomes that replication is called RNA and it generates proteins. In this paper, the tumor protein p53 is under consideration and it is analyzed critically. The tumor protein p53 dataset was first aligned and then analyzed on different tools of the MEME suite and Motif Finder. Every protein sequence contains a lot of residues. Fig. 3 shows all residues and their frequency.



A	Alanine	0.065	0.065
C	Cysteine	0.045	0.045
D	Aspartic acid	0.047	0.047
E	Glutamic acid	0.075	0.075
F	Phenylalanine	0.031	0.031
G	Glycine	0.059	0.059
H	Histidine	0.032	0.032
I	Isoleucine	0.029	0.030
K	Lysine	0.051	0.051
L	Leucine	0.089	0.089
M	Methionine	0.028	0.028
N	Asparagine	0.034	0.034
P	Proline	0.085	0.085
Q	Glutamine	0.037	0.038
R	Arginine	0.069	0.069
S	Serine	0.083	0.083
T	Threonine	0.051	0.051
V	Valine	0.056	0.056
W	Tryptophan	0.009	0.009
Y	Tyrosine	0.024	0.024

Fig. 4. Parameters for Motif detection.

B. *Result through MEME: Motif Detection*

Tumor protein sequence p53 is evaluated at MEME tool (Multiple Em for Motif Elicitation) that is a most famous tool for motif discovery and gained much attention in the field of bioinformatics. P53 based on probabilistic model and detects motifs by defining the probabilistic measures. p53 is the most important novel motif detection tool that takes the sequence as an input and finds the innovative motifs that are repeated patterns in the sequence. MEME divides these variable length patterns into unique un-gapped motifs. MEME work on some parameters to finds the unique motifs with length 166 to 1134. Table II shows all those parameters that are used to detect motifs in MEME. Based on these parameters, the motifs are detected through MEME tool some of these un-gapped motifs are shown in Fig. 4, 5 and 6, respectively at different alignment. Through MEME tool box the first alignment is just to repeat the motifs in a given sequence. The second alignment finds the motifs in the whole sequence. In order to determine the reminder sequence from the given dataset, the third alignment is performed so that all the left and the right possibility of motifs have been determined. This alignment further helps to detect the best motif sites in the Motif Enrichment Analysis (MEA). Between the width of 6 and 50, the motifs, MEME is analyzed the total three basic motifs that frequently presents among the whole sequence. In this experiment, it has been noted that the detection of motifs is usually present in the middle of the p53 protein sequence.

TABLE. II. SIMULATION PARAMETERS OF MEME

Parameters	*Description*
Sequences	*A set of 393 protein sequences of p53*
Length	*Between 166 and 1134*
Background	*Order-0 background generated by MEME according to given sequences*
Distribution	*One occurrence per sequence*
Motif width	*Between 6 wide and 50 wides*
Site	*EPLQVAHYRE**YWEYSIMC**ENKRTEQSVF* *EAYLI**YVCMKIIC**TGEELRVKES* *CSLQPSYSVL**FLGYLDMC**ABQERMRTYI*
Relative Entropy	*31.5*
Bayes Threshold	*10.2515*
p-value respectively	*6.29e-12* *9.71e-10* *2.77e-9*

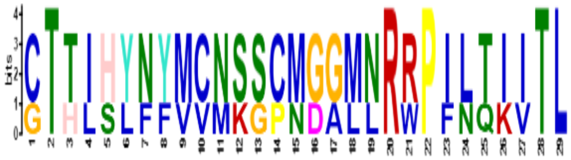


Fig. 4 First Alignment of Motif detection: for the sake of replication Scenario in order to determine the Motif discovery among the whole sequence.

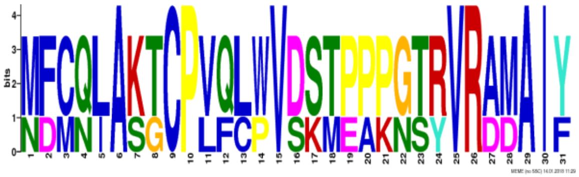


Fig. 5. Second alignment of Motif detection: for the sake of replication Scenario in order to determine the Motif discovery among the whole sequence.

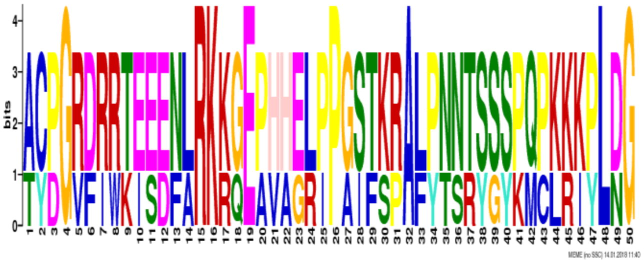
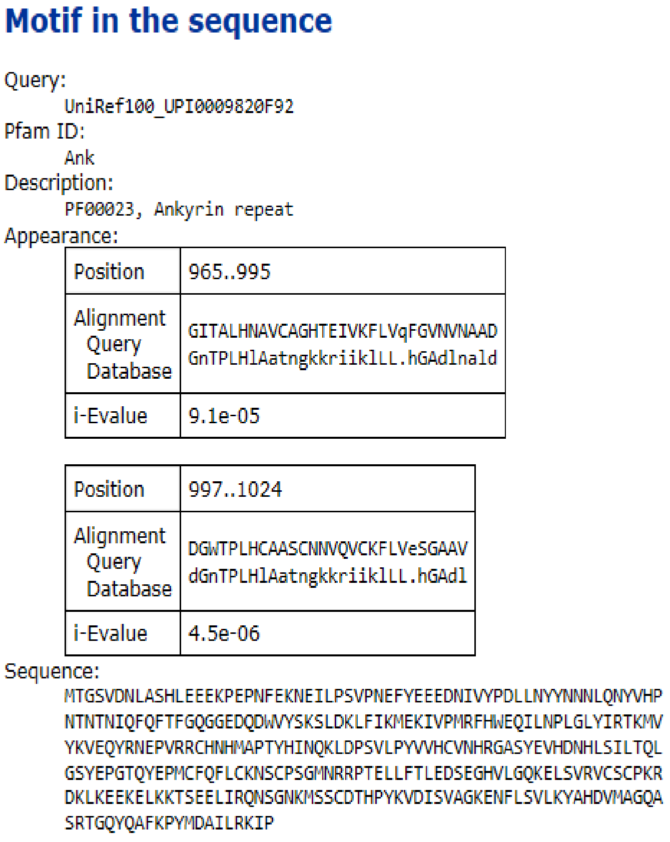


Fig. 6. Third alignment of Motif detection: for the sake of replication Scenario in order to determine the Motif discovery among the whole sequence.

VII. RESULT THROUGH MOTIF FINDER TOOL BOX: MOTIF DETECTION OF TUMOR PROTEIN

The tumor protein p53 is also analyzed by Motif Finder tool Box to check the multiple behaviors of this protein. The protein sequence is first aligned and then output motifs are detected. The resultant output of the Motif Finder is shown in Fig. 7.



Motif in the sequence

Query:
UniRef100_UPI0009820F92
Pfam ID:
Ank
Description:
PF00023, Ankyrin repeat
Appearance:

Position	965..995
Alignment Query Database	GITALHNAVCAGHTEIVKFLVqFGVNVNAAD GnTPLHlAatngkkriiklLL.hGAdlnald
i-Evalue	9.1e-05

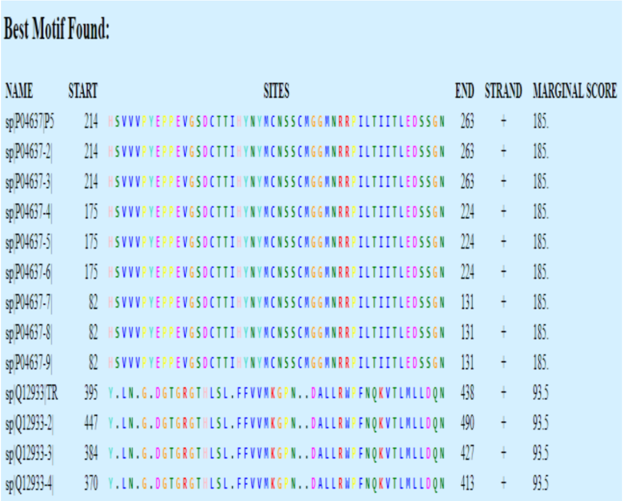
Position	997..1024
Alignment Query Database	DGWTPLHCAASCNNVQVCKFLVeSGAAV dGnTPLHlAatngkkriiklLL.hGAdl
i-Evalue	4.5e-06

Sequence:
```
MTGSVDNLASHLEEEKPEPNFEKNEILPSVPNEFYEEEDNIVYPDLLNYYNNNLQNYVHP
NTNTNIQFQFTFGQGGEDQDWVYSKSLDKLFIKMEKIVPMRFHWEQILNPLGLYIRTKMV
YKVEQYRNEPVRRCHNHMAPTYHINQKLDPSVLPYVVHCVNHRGASYEVHDNHLSILTQL
GSYEPGTQYEPMCFQFLCKNSCPSGMNRRPTELLFTLEDSEGHVLGQKELSVRVCSCPKR
DKLKEEKELKKTSEELIRQNSGNKMSSCDTHPYKVDISVAGKENFLSVLKYAHDVMAGQA
SRTGQYQAFKPYMDAILRKIP
```

Fig. 7. Motif detection through Motif finder.

A. *Result through GLAME2: for the Sake of Gapped Motif Detection*

To find out the gapped or rigid motifs in the tumor protein sequence p53 GLAME2 tool is used. The unique nature of this tool is that it finds the motifs of variable length. By GLAME2 the Best Motif is finding out in Fig. 8.



Best Motif Found:

NAME	START	SITES	END	STRAND	MARGINAL SCORE
sp|P04637|P5	214	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	263	+	185.
sp|P04637-2|	214	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	263	+	185.
sp|P04637-3|	214	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	263	+	185.
sp|P04637-4|	175	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	224	+	185.
sp|P04637-5|	175	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	224	+	185.
sp|P04637-6|	175	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	224	+	185.
sp|P04637-7|	82	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	131	+	185.
sp|P04637-8|	82	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	131	+	185.
sp|P04637-9|	82	HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN	131	+	185.
sp|Q12933|TR	395	Y.LN.G.DGTGRGTHLSL.FFVVMKGPN..DALLRWPFNQKVTLMLLDQN	438	+	93.5
sp|Q12933-2|	447	Y.LN.G.DGTGRGTHLSL.FFVVMKGPN..DALLRWPFNQKVTLMLLDQN	490	+	93.5
sp|Q12933-3|	384	Y.LN.G.DGTGRGTHLSL.FFVVMKGPN..DALLRWPFNQKVTLMLLDQN	427	+	93.5
sp|Q12933-4|	370	Y.LN.G.DGTGRGTHLSL.FFVVMKGPN..DALLRWPFNQKVTLMLLDQN	413	+	93.5

Fig. 8. Best Motif detection through GLAME2.

B. Gapped Motif Detection at different Alignment

In order to determine the gapped motifs, the three different alignments have been performed in this Fig. 9. There is an inverse relationship between the alignment and its relative score. As the number of alignment is increased the relative score is decreased. The reason is that the motif enrichment is decreased when same data sets are evaluated again and again for the same data sets.

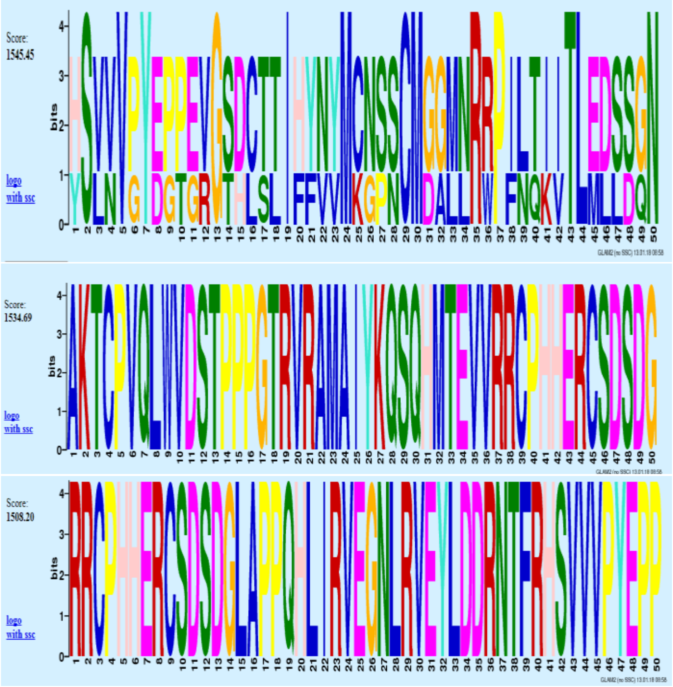
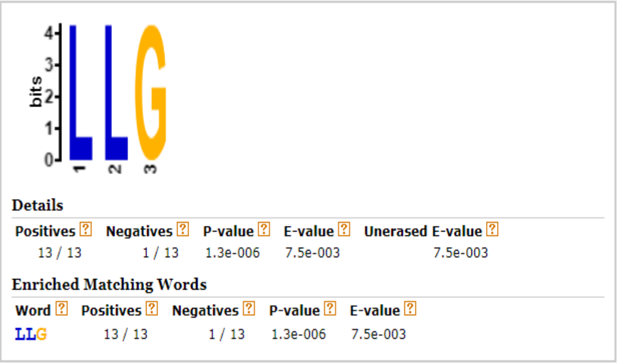


Fig. 9. Motif detection through GLAME2 with their relative score.

C. Result through DREME Tool Box: Un-gapped Motif Detection

In Fig. 10 and 11 un-gapped motifs are found by using DREME tool that discovers the short un-gapped motifs. The resultant output shows that there are only three motifs in the sequence that is un-gapped.



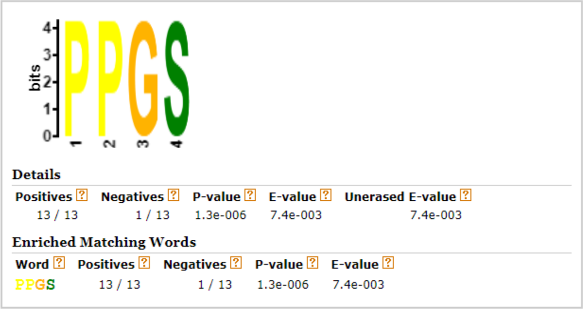
Details

Positives	Negatives	P-value	E-value	Unerased E-value
13 / 13	1 / 13	1.3e-006	7.5e-003	7.5e-003

Enriched Matching Words

Word	Positives	Negatives	P-value	E-value
LLG	13 / 13	1 / 13	1.3e-006	7.5e-003

Fig. 10. Motif detection through DREME.



Details

Positives	Negatives	P-value	E-value	Unerased E-value
13 / 13	1 / 13	1.3e-006	7.4e-003	7.4e-003

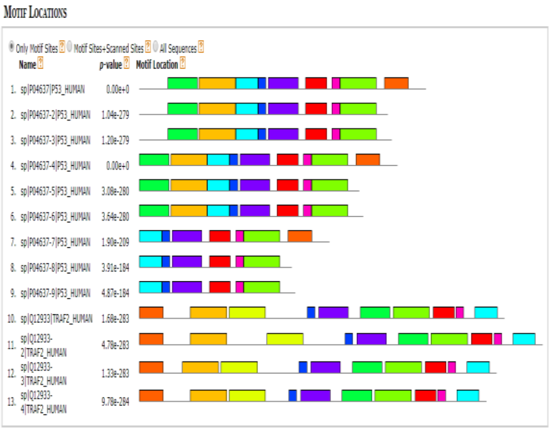
Enriched Matching Words

Word	Positives	Negatives	P-value	E-value
PPGS	13 / 13	1 / 13	1.3e-006	7.4e-003

Fig. 11. Motif detection through DREME.

D. Motif Locations in the Tumor Protein Sequence

Fig. 12 shows the motifs location in the tumor protein sequence. To find out the motifs matching factor in the protein sequences, different color scheme is used that represent it clearly. The un-matched sequence appears in red color while the matched sequences across the tumor protein are displayed with the blue color. In order to highlight the location of best motif sequence, the green color is displayed.



Motif Locations

● Only Motif Sites ○ Motif Sites+Scanned Sites ○ All Sequences

	Name	p-value	Motif Location
1.	sp|P04637|P53_HUMAN	0.00e+0	
2.	sp|P04637-2|P53_HUMAN	1.04e-279	
3.	sp|P04637-3|P53_HUMAN	1.20e-279	
4.	sp|P04637-4|P53_HUMAN	0.00e+0	
5.	sp|P04637-5|P53_HUMAN	3.08e-280	
6.	sp|P04637-6|P53_HUMAN	3.64e-280	
7.	sp|P04637-7|P53_HUMAN	1.90e-209	
8.	sp|P04637-8|P53_HUMAN	3.91e-184	
9.	sp|P04637-9|P53_HUMAN	4.87e-184	
10.	sp|Q12933|TRAF2_HUMAN	1.68e-283	
11.	sp|Q12933-2|TRAF2_HUMAN	4.78e-283	
12.	sp|Q12933-3|TRAF2_HUMAN	1.33e-283	
13.	sp|Q12933-4|TRAF2_HUMAN	9.78e-284	

Fig. 12. Motif location in tumor protein at different colors.

E. Motif Enrichment Analysis in Tumor Protein p53

Motif Enrichment Analysis (MEA) of tumor protein is done by using the Centrimo tool that selects those motifs that have a maximum number of repetitions. It takes the previous motifs that were detected by the MEME, Motif finder, GLAME2, DREME as input and applies the enrichment on it.

The enrichment is done through some specific parameters like a number of motifs, motifs width and the match score of the motifs. Fig. 13 shows the Enriched Motif Graph on the basis of parameters that are described in Table III. The sequence is extracted for the probability value 1.0 against the position of best sites in the sequence. Three sub-datasets are used that are commonly evaluated among all the above tools. These three sub-datasets in the form of motifs is described as WTPFHCAASC, WWWARLGD, and CHLAEVWCG.

TABLE. III. MOTIF ENRICHMENT PARAMETERS FOR THE ENRICHMENT ANALYSIS OF MEME, DREME, GLAME2 AND MOTIF FINDER TOOL BOX

Parameters	Description
Motifs	set of 10 motifs
Width	10 and 15
Average width	15
Background	Order -0 backgrounds were generated
Match Score	Sites considered where match score ≥ 5
Region E-value	E –value ≤ 10

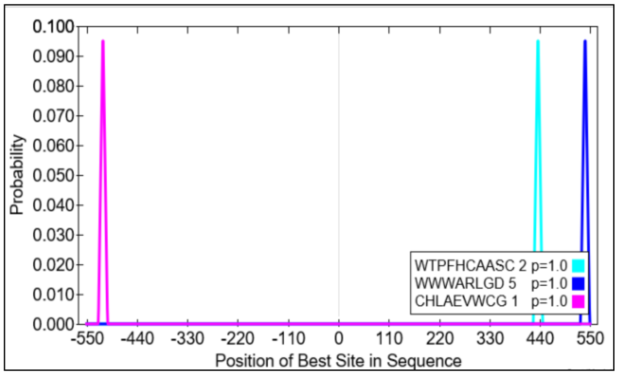


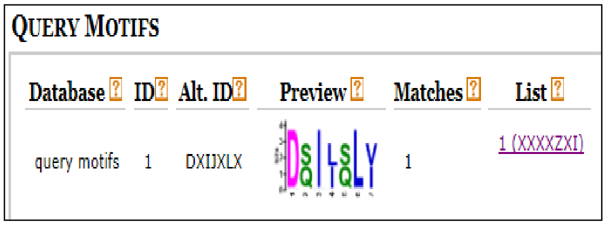
Fig. 13. Motifs Enrichment of Tumor Protein p53

F. Comparison of Tumor Protein p53 by TOMTOM TOOL BOX

The comparison is done of tumor protein p53 motifs with other protein motifs by using TOMTOM that is the best tool for motif comparison. The matching parameters for the TOMTOM comparison are given in Table IV. Fig. 14 shows the matched motifs of protein structure. Only one hit was found in the TOMTOM database that has been displayed below.

TABLE. IV. MOTIF COMPARISON PARAMETERS OF MOTIF DETECTION AMONG OTHER SPECIES

Comparison Parameters	
p-value	2.04e-04
e-value	2.04e-04
q-value	2.04e-04
Overlap	7
Offset	0
Orientation	Normal



QUERY MOTIFS

Database ?	ID ?	Alt. ID ?	Preview ?	Matches ?	List ?
query motifs	1	DXIJXLX		1	1 (XXXXZXI)

Fig. 14. Motifs comparison by TOMTOM

VIII. COMPARISON OF P53 WITH OTHER SPECIES

A. Comparison of Tumor Protein p53 Homo Species with Other Species with same Tumor Protein p53

Tumor protein p53 exists in multiple species. In Table IV the comparison of homo tumor protein p53 is done with other species that contain same tumor protein. The comparison is done on the previously detected motifs by the MEME, GLAME2, DREME and Motif Finder and number of matched motifs is found that is mentioned in Table V.

TABLE. V. P53 MOTIF COMPARISON PARAMETERS WITH OTHER SPECIES

Knowledgebase	No of Matches
MOUSE	3
JASPER	4
Homo Sapiens	8
FLY (combined drosophila database)	2
CIS-BP Single species	1
Prokaryote DNA (CoLlecTF (bacterial	5
Ray 2013 all species (DNA Encoded)	3
YEASTRACT	3
Swiss Regulon e coli	9
DAP Motifs	0
Vertebrates	1
Malaria	9

Motifs exist in the homo species with different shapes formats and length at a different location. Due to the diversity of nature of motifs, the discovery of motifs is challenging task. Some motifs are site-based and some are residue-based. In Tables VI and VII, the motif detection is done by using multiple tools against the same tumor protein sequences. Firstly, the site based motif is detected by using MEME, GLAME2, HHMOTIF and SLIM Finder than residue based motifs is detected by using same tools and protein sequences. The results show that MEME tool is most appropriate to find the site and residue-based motifs in the tumor proteins respectively. The two parameters are used for this propose like recall and precision. Recall states that the relevant motifs among the retrieved motifs and the precision states that the relevant motif that should be retrieved and the collection of precision and recall is F1-measure.

TABLE. VI. SITE BASED COMPARISON OF MOTIF

Site-based

Tools	Recall	Precision	F1
MEME	0.236	0.564	0.333
GLAM2	0.249	0.099	0.142
HH-MOTIF	0.413	0.164	0.235
SLIM Finder	0.272	0.389	0.320

TABLE. VII. RESIDUE BASED COMPARISON OF MOTIF

Residue-based

Tools	Recall	Precision	F1
MEME	0.210	0.420	0.280
GLAM2	0.219	0.061	0.095
HH-MOTIF	0.380	0.073	0.123
SLIM Finder	0.203	0.350	0.257

In Fig. 15 the accuracy comparison of different tools that are available for motifs detection is done and the performance of each tool is measured in the graphical format. Two parameters are under consideration, the first one is an F1 site that shows the site-based motifs in the sequence and the second one is the F1-residue that shows the residue-based motifs in the protein sequences. MEME tool is the only tool that finds the site-based and residue-based motifs with a high score and stood first in all of the other motif detection tools that are GLAM2, HHMOTIF, and SLIM Finder for the hundreds number of iterations.

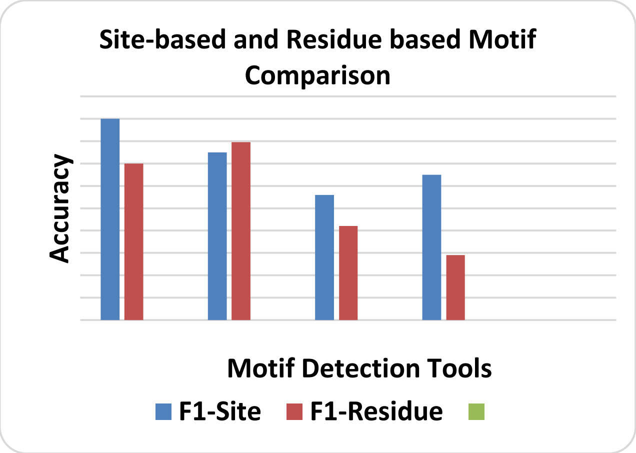


Fig. 15. Performance as measured By F1 of MEME and other tested de novo slim search methods.

IX. CONCLUSION

Big data became an active research from the last few years due to its immeasurable range of applications. A lot of research has been done at the motif detection in the field of big data and bioinformatics, but still more attention is required in this field. Motif detection is some of the highly focused topics among the researchers in the big data research community. These Motifs are useful for various time-series and data mining tasks. The relation between DNA and protein is a key motivating force. Binding of protein-sites and the specially targeted proteins are two important moves to understand the concept of biological activities. The combined effort of p53 and p63 in some differential composition of DNA-binding sites may contribute to distinct functions of these proteins homologs in some different species. The p53 tumor suppressor is implicated in cell cycle control, DNA repair, explicative sequence and programmed cell death. Inactivation of the p53 contributes to the wide range of human tumors; including Glial neoplasm's. Due to its lot of benefits and features its need to analyze the tumor protein more critically. In this paper, the proposed algorithm is introduced to detect the motifs of different lengths with gaps and without gaps through p53 Motif Detection Algorithm. The sequence analysis of the tumor protein p53 is performed by using MEME tool. MEME tool is the only tool that finds the site-based and residue-based motifs with a high score and stood first in all of the other motif detection tools that are GLAM2, HHMOTIF, and SLIM Finder.

In this paper we have originated the problem of detecting motifs in the tumor proteins p53 that is depicted as *"the guardian of the genome"*, referring to its role in persevering stability by preventing genome mutation and have offered a universal scheme for it. The P53MDA is purposed to detect the motifs in the tumor protein. Our formulation of the problem provides for a rigorous measure of the best fit between a given pattern and an example.

Up to our best knowledge, this work is firstly done on the basis of both data types format. To select the discovery mode discriminative then the number of protein are aligned is some order otherwise random order applies. The novelty of this work is not only the detection of motif inside the protein but also provides the detailed comparison among different available tools. The possible future direction is that to find out the best motifs and its correct alignment in a well-disciplined manner, in this way the diseases that are associated the DNA and Genome structure is easily trace out.

REFERENCES

[1] Maass, W., Parsons, J., Purao, S., Rosales, A., Storey, V. C., & Woo, C. C. (2017). Big Data and Theory. Encyclopedia of Big Data, 1-5.

[2] Rai, A., Pradhan, P., Nagraj, J., Lohitesh, K., Chowdhury, R., & Jalan, S. (2017). Understanding cancer complexome using networks, spectral graph theory and multilayer framework. Scientific reports, 7, 41676.

[3] Saha, T. K., Katebi, A., Dhifli, W., & Al Hasan, M. (2017). Discovery of Functional Motifs from the Interface Region of Oligomeric Proteins using Frequent Subgraph Mining. IEEE/ACM transactions on computational biology and bioinformatics.

[4] Lipper, C. H., Karmi, O., Sohn, Y. S., Darash-Yahana, M., Lammert, H., Song, L., ... & Jennings, P. A. (2018). Structure of the human monomeric NEET protein MiNT and its role in regulating iron and reactive oxygen species in cancer cells. Proceedings of the National Academy of Sciences, 115(2), 272-277.

[5] Ma, W., Noble, W. S., & Bailey, T. L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nature protocols, 9(6), 1428-1450.

[6] Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831-838.

[7] Fan, C., Xiao, F., Madsen, H., & Wang, D. (2015). Temporal knowledge discovery in big BAS data for building energy management. Energy and Buildings, 109, 75-89.

[8] Tran, N. T. L., & Huang, C. H. (2017). Cloud-based MOTIFSIM: Detecting Similarity in Large DNA Motif Data Sets. Journal of Computational Biology, 24(5), 450-459.

[9] Schröter, M., Paulsen, O., & Bullmore, E. T. (2017). Micro-connectomics: probing the organization of neuronal networks at the cellular scale. Nature Reviews Neuroscience, 18(3), 131-146.

[10] Yang, J., Jiang, B., Li, B., Tian, K., & Lv, Z. (2017). A fast image retrieval method designed for network big data. IEEE Transactions on Industrial Informatics.

[11] Chen, D., Jiang, S., Ma, X., & Li, F. (2017). TFBSbank: a platform to dissect the big data of protein–DNA interaction in human and model species. Nucleic acids research, 45(D1), D151-D157.

[12] Fan, C., Xiao, F., Madsen, H., & Wang, D. (2015). Temporal knowledge discovery in big BAS data for building energy management. Energy and Buildings, 109, 75-89.

[13] Wang, Y., Wang, H., & Chang, S. (2018). A weighted higher-order network analysis of fine particulate matter (PM2. 5) transport in Yangtze River Delta. Physica A: Statistical Mechanics and its Applications.

[14] Nagaraj, K., Sharvani, G. S., & Sridhar, A. (2018). Emerging trend of big data analytics in bioinformatics: a literature review. International Journal of Bioinformatics Research and Applications, 14(1-2), 144-205.

[15] de França, F. O., Goya, D., & Penteado, C. C. (2018, January). Analysis of the Twitter Interactions during the Impeachment of Brazilian President. In Proceedings of the 51st Hawaii International Conference on System Sciences.

[16] Hoh, J., Jin, S., Parrado, T., Edington, J., Levine, A. J., & Ott, J. (2002). The p53MH algorithm and its application in detecting p53-responsive genes. Proceedings of the National Academy of Sciences, 99(13), 8467-8472.

[17] Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., ... & Liu, J. (2006). A global map of p53 transcription-factor binding sites in the human genome. Cell, 124(1), 207-219.

[18] Li, W., Meyer, C. A., & Liu, X. S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics, 21(suppl_1), i274-i282.

[19] Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., ... & Attardi, L. D. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell, 142(3), 409-419.

[20] Zhang, Y., Hu, Y., Wang, J. L., Yao, H., Wang, H., Liang, L. & Xu, J. (2017). Proteomic identification of ERP29 as a key chemoresistant factor activated by the aggregating p53 mutant Arg282Trp. Oncogene, 36(39), 5473.

[21] Higgins, S. P., Tang, Y., Higgins, C. E., Mian, B., Zhang, W., Czekay, R. P., ... & Higgins, P. J. (2018). TGF-β1/p53 signaling in renal fibrogenesis. Cellular signalling, 43, 1-10.

[22] Chowdhury, K., Kumar, S., Sharma, T., Sharma, A., Bhagat, M., Kamai, A., ... & Mandal, C. C. (2018). Presence of a consensus DNA motif at nearby DNA sequence of the mutation susceptible CG nucleotides. Gene, 639, 85-95.