A study of P53 gene and its regulatory genes network

Jayanta Kumar Das
Applied Statistics Unit
Indian Statistical Institute
Kolkata-700108, India
dasjayantakumar89@gmail.com

Suvankar Ghosh
Department of Biosciences and Bioengineering
IIT Guwahati
Guwahati-781039, India
svnkrgsh@gmail.com

Ranjeet Kumar Rout
Amity School of Engineering and Tecnology,
Amity University,
Noida-201313, India
ranjeetkumarrout@gmail.com

Pabitra Pal Choudhury
Applied Statistics Unit
Indian Statistical Institute
Kolkata-700108, India
pabitrapalchoudhury@gmail.com

***Abstract*-Gene network analysis and the mutation detection is always a cumbersome process both for biologists and non-biologist people. So an innovative approach is essential to counteract this problem. In this manuscript, we want to uniquely characterize the four nucleotides (A, T, C & G) in each interacting genes in the P53 gene regulatory network and 2D graphical representation of genes using their chemical properties (purine and pyrimidine). This representation will help to analyze and identify the mutated gene in the gene regulatory network both for the computational biologist as well as the mathematician. Furthermore, we look for the similarity analysis of the amino acids sequences of the associated genes of TP53 genes network and it is found that the amino acids sequences are significantly dissimilar. We find the three highest conserved protein blocks of length 4 amino acids that may play some significant role for their network functioning.**

Keywords- P53; Gene network; 3D graph; DNA walk; Protein Sequences; Pattern.

I. INTRODUCTION

The TP53 gene provides instructions for making a protein called tumor protein P53 (or TP53). This protein plays most important role in preventing tumour development [1], acts as tumour suppressor [2, 3, 4, 5, 6]. It responds to a range of potentially oncogenic stresses by activating protective mechanisms [7], most notably cell cycle arrest and apoptosis. Its significance as a tumour suppressor is reflected by its high rate of mutation in human cancer [1], with more than 50% of adult human tumours bearing inactivating mutations or deletions in the **T**P53gene.

The inactivation of the TP53 gene is due to the varieties of small mutations: missense and nonsense mutations or insertions/deletions of several nucleotides etc. [8][9], which lead to either expression of a mutant protein (90% of cases) or absence of a protein (10% of cases). TP53 is found to be involved in around of 50% of all cancers cells (tumour) where its different types of mutations are observed [8].

The pathway of p53 gene is the collection of genes and its interacting links (forming a network) that are targeted to respond to a variety of intrinsic and extrinsic signals that impact upon the mechanisms that monitor DNA replication, chromosome segregation and cell division [10]. In response to a signal, the p53 protein is activated in a specific manner by post-translational modifications, and this leads to either cell cycle arrest, a program that induces cell senescence or cellular apoptosis [11]. In this way, a variety of intrinsic or extrinsic stresses that would result in a loss of fidelity in DNA replication, genome stability, cell cycle progression or faithful chromosome segregation can be accommodated or, alternatively, the clone of cells with these defects eliminated from the body.

In this manuscript, we apply a 3D graphical representation (discussed in [12, 13]) of DNA primary sequences for unique characterization of P53 gene and its interacting genes. This will help us to identify the mutated nucleotide in the gene. We also 2D plot the each gene using their chemical characteristics (using purine and pyrimidine class) named random DNA walk [13], from which mutated position can be confined. This experimental work is extremely helpful and this can be useful to characterize any gene regulatory network. Apart from this, we also go for in depth understanding of protein sequences associated genes for P53 genes regulatory network [14, 15, 16].

II. MATERIALS

P53 gene and its corresponding interacting genes are collected from the STRING server shown in Fig. 1 [10]. P53 gene is associated with the 11 genes which are shown in TABLE I. The DNA sequences (partial) of each interacting genes are also shown in TABLE I.

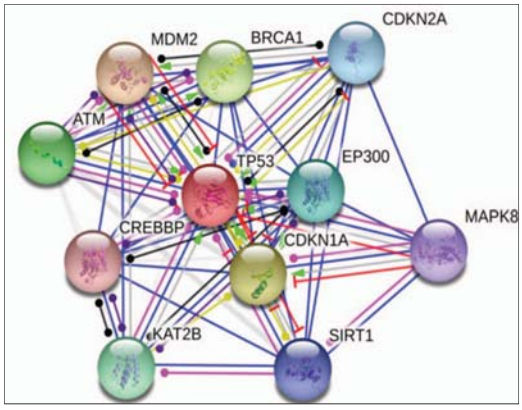

Fig. 1. TP53 and its interactive genes network

III. METHODOLOGY AND RESULTS

A. 3D Unique Representation of Nucleotides

3D Unique representations of each nucleotide are as follows:

$\left(-\mathbf{1},\mathbf{0},\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{i}}\right)\rightarrow \boldsymbol{A}$, $\left(\mathbf{0},-\mathbf{1},\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{i}}\right)\rightarrow \boldsymbol{T}$,

$\left(\mathbf{1},\mathbf{0},\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{i}}\right)\rightarrow \boldsymbol{G}$, $\left(\mathbf{0},\mathbf{1},\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{i}}\right)\rightarrow \boldsymbol{C}$

In this notation, $\boldsymbol{i}$ represent the position of that particular nucleotide in the DNA sequence; from this we can get unique sets of points for each nucleotide of every DNA sequence. Interested person can go though at [13] for more details on 3D unique representation of DNA sequence which is shown in TABLE II.

TABLE I. DNA SEQUENCES OF TP53 AND ITS INTERACTING GENES (PARTIAL)

Gene Name	DNA Sequences
TP53	ATGGAGGAGCCGCAGTCAGATCCTAGCGTCGAG CCCCCTCTGAGTCAGGAAA.....
ATM	ATGAGTCTAGTACTTAATGATCTGCTTATCTGCT GCCGTCAACTAGAACATGA.....
BRCA1	ATGGATTTATCTGCTCTTCGCGTTGAAGAAGTAC AAAATGTCATTAATGCTAT.....
CDKN1A	ATGTCAGAACCGGCTGGGGATGTCCGTCAGAAC CCATGCGGCAGCAAGGCCT.....
CDKN2A	ATGGAGCCGGCGGCGGGGAGCAGCATGGAGCCT TCGGCTGACTGGCTGGCC.....
CREBBP	ATGGCTGAGAACTTGCTGGACGGACCGCCCAAC CCCAAAAGAGCCAAACTCAG.....
EP300	ATGGCCGAGAATGTGGTGGAACCGGGGCCGCCT TCAGCCAAGCGGCCTAAAC.....
KAT2B	ATGTCCGAGGCTGGCGGGGCCGGGCCGGGCGG CTGCGGGGCAGGAGCCGGG.....
MAPK8	ATGAGCAGAAGCAAGCGTGACAACAATTTTTATA GTGTAGAGATTGGAGATTC.....
MDM2	ATGTGCAATACCAACATGTCTGTACCTACTGATG GTGCTGTAACCACCTCACA.....
SIRT1	ATGTTTGATATTGAATATTTCAGAAAAGATCCAA GACCATTCTTCAAGTTTGCA.....

Example: Take the first 10 nucleotides of P53 gene which is ATGGAGGAGC and their corresponding co-ordinates as follows:

$$\left\{\begin{matrix}(-1,0,0),\left(0,-1,\frac{1}{2}\right),\left(1,0,\frac{2}{3}\right),\left(1,0,\frac{3}{4}\right),\left(-1,0,\frac{4}{5}\right),\\ \left(1,0,\frac{5}{6}\right),\left(1,0,\frac{6}{7}\right),\left(-1,0,\frac{7}{8}\right),\left(1,0,\frac{8}{9}\right),\left(0,1,\frac{9}{10}\right)\end{matrix}\right\}$$

Now, we calculate the average of each nucleotide as shown below.

$$A = average\left\{(-1,0,0),\left(-1,0,\frac{4}{5}\right),\left(-1,0,\frac{7}{8}\right)\right\}$$

$$A = (-1,0,0.55833)$$

$$G = average\left\{\begin{matrix}\left(1,0,\frac{2}{3}\right),\left(1,0,\frac{3}{4}\right),\left(1,0,\frac{5}{6}\right),\left(1,0,\frac{6}{7}\right),\\ \left(1,0,\frac{8}{9}\right)\end{matrix}\right\}$$

$$G = (1,0,0.799207)$$

$$T = average\left\{\left(0,-1,\frac{1}{2}\right)\right\}$$

$$T = (0,-1,0.5)$$

$$C = average\left\{\left(0,1,\frac{9}{10}\right)\right\}$$

$$C = (0,1,0.9)$$

TABLE II. 3D REPRESENTATION OF DNA SEQUENCES TP53 AND ITS INTERACTING GENES

Gene Name	Nucleotide	Coordinate		
		X	Y	Z
P53	A	-1	0	0.991282622647492
	T	0	-1	0.993726105088182
	G	1	0	0.993028243528498
	C	0	1	0.995451968351100
ATM	A	-1	0	0.999073375322011
	T	0	-1	0.998703451480067
	G	1	0	0.998890971101964
	C	0	1	0.999144191701159
BRCA1	A	-1	0	0.998303842760123
	T	0	-1	0.997976776489626
	G	1	0	0.998471956383968
	C	0	1	0.998781921645161
CDKN1A	A	-1	0	0.978528402183162
	T	0	-1	0.983609022190974
	G	1	0	0.988518779949124
	C	0	1	0.989976482426877
CDKN2A	A	-1	0	0.974611872179305
	T	0	-1	0.985097269188758
	G	1	0	0.985629481012738
	C	0	1	0.990664514126318
CREBBP	A	-1	0	0.998420829092023
	T	0	-1	0.998474458021157
	G	1	0	0.998667052234219
	C	0	1	0.999040031279974
EP300	A	-1	0	0.999060874147456
	T	0	-1	0.998349058522068
	G	1	0	0.998257940587574
	C	0	1	0.998931478177949
KAT2B	A	-1	0	0.997314850238713
	T	0	-1	0.997250043091072
	G	1	0	0.995297212267122
	C	0	1	0.996550389139235
MAPK8	A	-1	0	0.992060158498567
	T	0	-1	0.994803262078009
	G	1	0	0.993960180093589
	C	0	1	0.995852111464939
MDM2	A	-1	0	0.993614827858114
	T	0	-1	0.993890221188402
	G	1	0	0.994954571563894
	C	0	1	0.994615647473085
SIRT1	A	-1	0	0.993440770269773
	T	0	-1	0.992823359253081
	G	1	0	0.995474867240593
	C	0	1	0.996407510030051

Suppose there is a mutation in 7th position where **G is** replaced by **C** i.e. new sequence is ATGGAG**C**AGC and their corresponding co-ordinates as follows:

$$\left\{\begin{matrix}(-1,0,0),\left(0,-1,\frac{1}{2}\right),\left(1,0,\frac{2}{3}\right),\left(1,0,\frac{3}{4}\right),\left(-1,0,\frac{4}{5}\right),\\ \left(1,0,\frac{5}{6}\right),\left(0,1,\frac{6}{7}\right),\left(-1,0,\frac{7}{8}\right),\left(1,0,\frac{8}{9}\right),\left(0,1,\frac{9}{10}\right)\end{matrix}\right\}$$

$$A = average\left\{(-1,0,0),\left(-1,0,\frac{4}{5}\right),\left(-1,0,\frac{7}{8}\right)\right\}$$

$A = (-1, 0, 0.55833)$

$G = average \left\{\left(1, 0, \frac{2}{3}\right), \left(1, 0, \frac{3}{4}\right), \left(1, 0, \frac{5}{6}\right), \left(1, 0, \frac{8}{9}\right)\right\}$

$G = (\mathbf{1, 0, 0.784722})$

$T = average \left\{\left(0, -1, \frac{1}{2}\right)\right\}$

$T = (0, -1, 0.5)$

$C = average \left\{\left(0, 1, \frac{6}{7}\right), \left(0, 1, \frac{9}{10}\right)\right\}$

$C = (\mathbf{0, 1, 0.878571})$

So, it can be observed that average C and G values getting changed along z-axis that shows the indication of mutation in the sequence.

B. 2D Representation of each DNA and its Position Identification

A DNA sequence is composed of the different combinations of four bases of nucleotides: {A, C, G, T}. These four bases are *A*, *C*, *G* and *T* can be divided into different classes. One of the important class is Purine = {*A*, *G*} and Pyrimidine = {*C*, *T*}.

We take **R = -1 (purine)** and **Y = +1 (pyrimidine).**

2D graphical representation of P53 and its interacting genes are shown here. Position of the genetic mutation in these genes can be detected via those plots.

For an example the sequence ATGGAGGAGC is changed to ATGGAGCAGC and the corresponding 2D Graph is shown in Fig. 2; where the sequence ATGGAGGAGC and ATGGAG**C**AGC are the original and mutated sequences respectively.

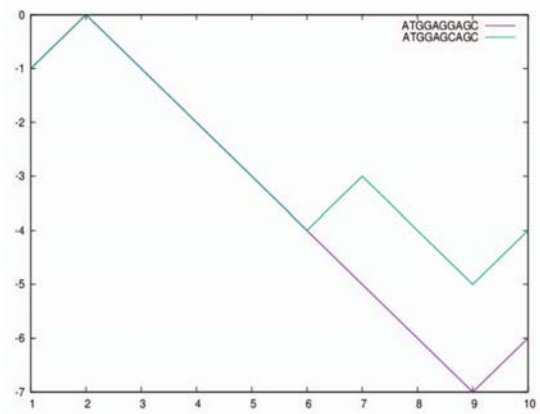

Fig. 2. Mutation position detection of a mutated sequence (aqua marine) with respect to its original sequence (blue violet)

From the above Fig. 2, it can be easily observed that at 7th position of that graph (aqua marine color) is shifted upward with +2 value and it continues that way, so we can say that the gene sequence is mutated at 7th position and if there is any further mutation that graph is again shifted either +2 or -2 position. Mutation position can be detected by this position shifting of these gene sequences.

Random DNA walk of P53 and its interacting genes are given below Fig. 3. (a-k). If we look into the 2D graph of DNA walk, we can see the huge variations among the graphs for different genes that signify position specific they are highly dissimilar.

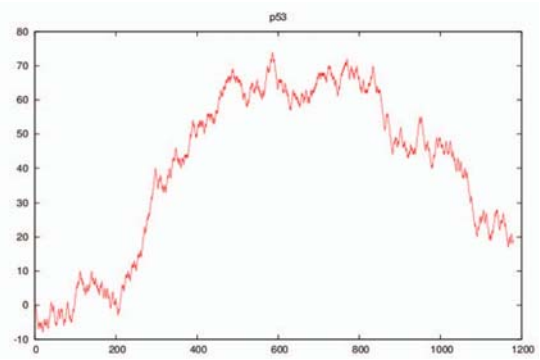

Fig. 3. (a) Random DNA walk of P53 gene

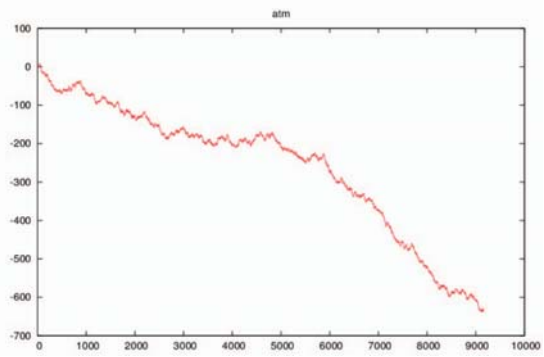

Fig. 3. (b) Random DNA walk of ATM gene

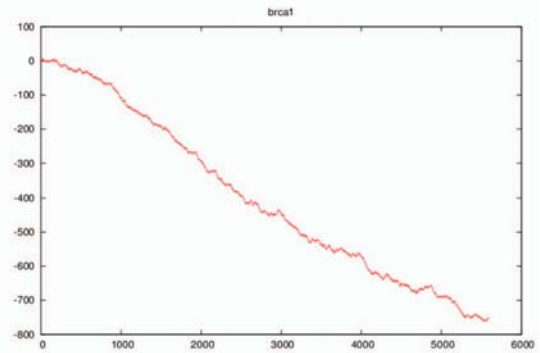

Fig. 3. (c) Random DNA walk of BRCA1 gene

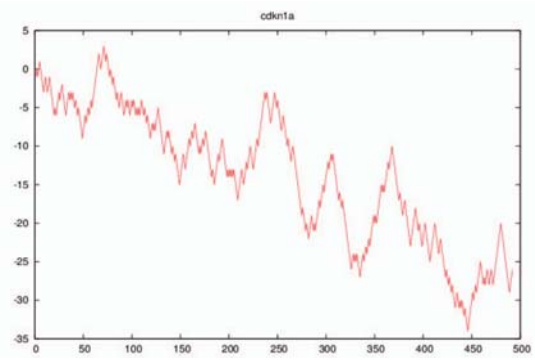

Fig 3. (d) Random DNA walk of CDKN1A gene

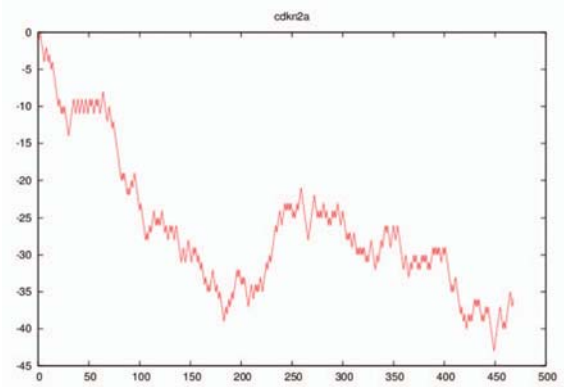

Fig. 3. (e) Random DNA walk of CDKN2A gene

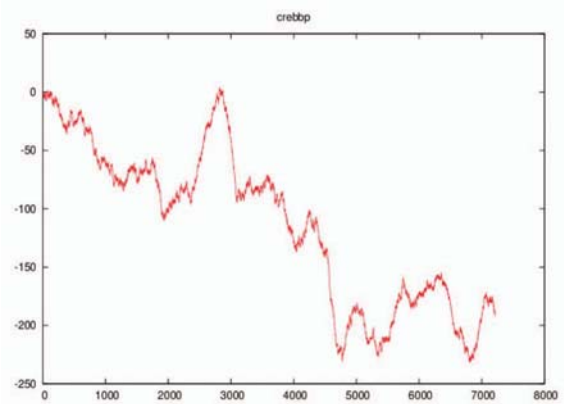

Fig. 3. (f) Random DNA walk of CREBBP gene

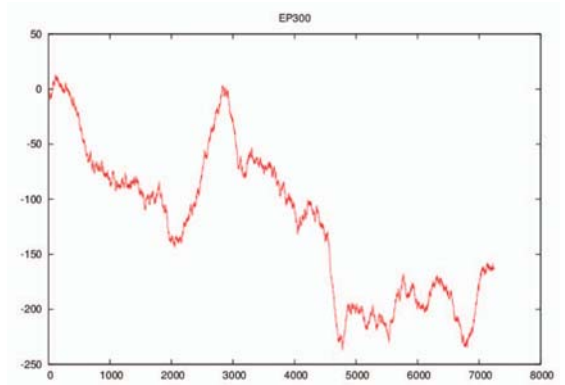

Fig. 3. (g) Random DNA walk of EP300 gene

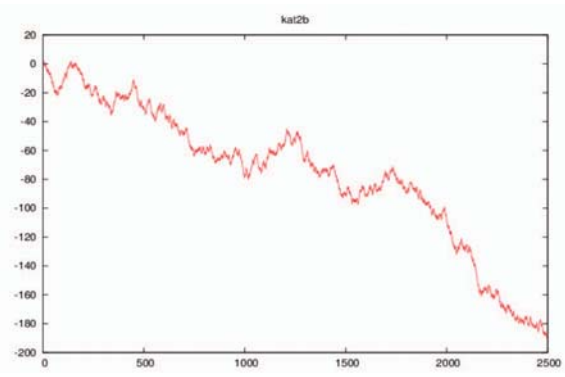

Fig. 3. (h) Random DNA walk of KAT2B gene

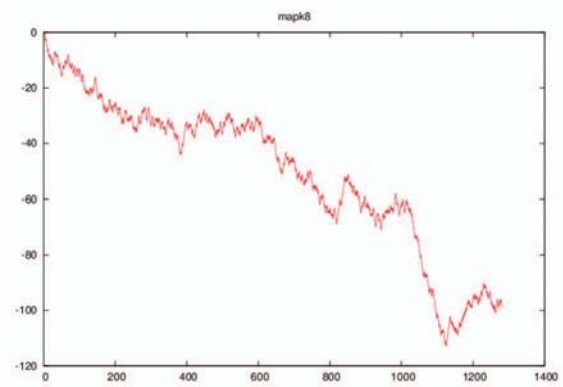

Fig. 3. (i) Random DNA walk of MAPK8 gene

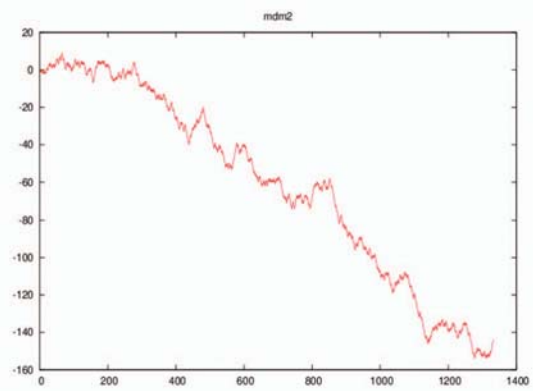

Fig. 3. (j) Random DNA walk of MDM2 gene

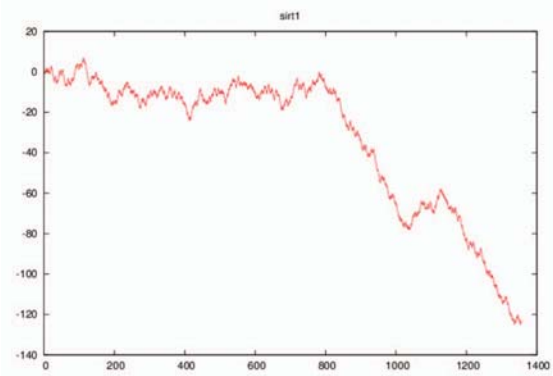

Fig. 3. (k) Random DNA walk of SIRT1 gene

C. Comparisons of Amino Acids Sequences

Protein sequences (or amino acids sequences) of the corresponding 11 genes of P53 genes regulatory network are collected from uniport data base (*http://www.uniprot.or)*. All the proteins with the unique accession number are shown in TABLE III. Based on multiple sequences alignment using Clustalw, it is observed sequences are approximately 35% similar in average and the highest similarity we find in between the pair sequence (3, 5). Further, analysis with chemical properties of amino acids in group wise as discussed in [14, 15], the similarity increased by 5% only as shown in TABLE IV. The pair sequences (6, 5), (1, 5) and (3, 5) are highly similar. It is also observed

that there are three patterns/blocks of length 4 amino acids which are chemically 100% conserved across all the sequences of P53 genes network. The patterns are "2244", "4414" and "4442" where numbers are based on chemical groups discussed in [14] (TABLE V). The location of each pattern is also shown in TABLE VI for all the sequences. In some cases, we find more than one similar pattern in different locations; we took only the first one and its position. These conserved patterns, based on chemical properties of amino acids, may play the significant role for doing their network functioning as in earlier we have found some block specific functional role when applied the same methodology for other different protein families [14, 16].

TABLE III. TP53 AND ITS ASSOCIATED GENES AND CORRESPONDING PROTEIN ACCESSION NUMBERS

Serial Numbers	Genes	Protein Accession Numbers
1	TP53	P04637
2	ATM	Q13315
3	BRCA1	P38398
4	CDKN1A	P38936
5	CDKN2A	P42771
6	CREBBP	Q92793
7	EP300	Q09472
8	KAT2B	Q92831
9	MAPK8	P45983
10	MDM2	Q00987
11	SIRT1	Q96EB6

TABLE IV. SIMILARITY PERCENTAGE (PAIRS-WISE) AMONG THE AMINO ACIDS SEQUENCES OF TP53 GENES NETWORK BASED ON THE EIGHT CHEMICAL GROUPS OF AMINO ACIDS [14]

Seq. Vs Seq.	1	2	3	4	5	6	7	8	9	10	11
1	0	23	20	17	43	23	25	22	20	22	25
2	23	0	24	17	38	36	38	13	9	26	23
3	20	24	0	25	43	29	32	29	25	13	28
4	17	17	25	0	30	33	35	13	16	29	17
5	43	38	43	30	0	54	55	34	32	52	30
6	23	36	29	33	54	0	5	32	32	36	34
7	25	38	32	35	55	5	0	34	34	39	36
8	22	13	29	13	34	32	34	0	7	34	19
9	20	9	25	16	32	32	34	7	0	28	18
10	22	26	13	29	52	36	39	34	28	0	26
11	25	23	28	17	30	34	36	19	18	26	0

TABLE V. AMINO ACIDS AND THEIR CHEMICAL GROUPS

Group Nos.	Chemical Nature	Amino Acids
1/G1	Acidic	Aspartate (D), Glutamate (E)
2/G2	Basic	Arginine (R), Histidine (H), Lysine (K)
3/G3	Aromatic side chain	Tyrosine (Y), Phenylalanine (F), Tryptophan (W)
4/G4	Aliphatic side chain	Isoleucine (I), Leucine (L), Valine (V), Alanine (A), Glycine (G)
5/G5	Cyclic	Proline (P)
6/G6	Sulfur containing	Methionine (M), Cysteine (C)
7/G7	Hydroxyl	Serine (S), Threonine (T)
	containing	
8/G8	Acidic amide	Glutamine (Q), Asparagine (N)

TABLE VI. THREE CONSERVED PATTERNS OF LENGTH 4 AMONG THE AMINO ACIDS SEQUENCES OF P53 NETWORK

	Patterns		
Seq. Nos.	"2244"	"4414"	"4442"
1	305-RRII	347-IIDI	264-IIIR
2	294-RRII	46-IIDI	63-IIIR
3	135-RRII	409-IIDI	312-IIIR
4	155-RRII	5-IIDI	90-IIIR
5	98-RRII	31-IIDI	19-IIIR
6	351-RRII	5-IIDI	215-IIIR
7	291-RRII	58-IIDI	197-IIIR
8	490-RRII	165-IIDI	75-IIIR
9	221-RRII	303-IIDI	52-IIIR
10	162-RRII	82-IIDI	33-IIIR
11	254-RRII	98-IIDI	62-IIIR

IV. DISCUSSION AND CONCLUSION

Here, the proposed methodology as discussed in [13] is quite simple and we apply this logic for mutation specific unique representation of genes of P53 genes regulatory network. Using this method, we can easily detect the occurrence of mutation because if there is a change of any nucleotide in DNA sequence, the value of Z co-ordinate will be altered. We also able to capture for finding which nucleotide is mutated from their 2D graph representation (DNA Walk). Further, according to analysis with the protein sequences of the corresponding genes of TP53 genes in the regulatory network, it is observed that they don't have any significant common signature with regards to the pattern of amino acids even if we search based on the transformed eight chemical group specific amino acids patterns and similarity [14,15]. We find only 4 length conserved patterns of amino acids across the sequences. These blocks/patterns might have some roles for the associations of P53 gene with its associated genes in P53 network.

REFERENCES

[1] S. Surget , M.P. Khoury, J.C. Bourdon. "Uncovering the role of p53 splice variants in human malignancy: a clinical perspective," Onco Targets and Therapy, vol-7, p:57-68. 2013 doi:10.2147/OTT.S53876.PMID 24379683.

[2] G. Matlashewski, P. Lamb, D. Pim, J. Peacock, L. Crawford, S. Benchimol, "Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene," EMBO J, 3 (13):3257-3262. PMC 557846, 1984.

[3] M. Isobe, B. S. Emanuel, D. Givol, M. Oren, C.M. Croce, "Localization of gene for human p53 tumour antigen to band 17p13," Nature, 320 (6057):84-85. doi:10.1038/320084a0. PMID 3456488, 1986.

[4] S. E. Kern, K. W. Kinzler, A. Bruskin, D. Jarosz, P. Friedman, C. Prives, "Identification of p53 as a sequence-specific DNA-binding protein," Science, 252(5013):1708-11. doi:10.1126/science.2047879. PMID 2047879, 1991

[5] O.W. McBride, D. Merry, D. Givol, "The gene for human p53 cellular tumor antigen is located on chromosome 17 short arm (17p13)," Proc. Natl. Acad. Sci. U.S.A. 83 (1): 130–134. doi:10.1073/pnas.83.1.130. PMC 322805.PMID 3001719, 1986.

[6] J.C. Bourdon, K. Fernandes, F. M. Zmijewski , G. Liu, A. Diot, D. P. Xirodimas, "p53 isoforms can regulate p53 transcriptional activity," Genes & Development ,19 (18): 2122-2137. doi:10.1101/gad.1339905. PMID 16131611, 2005

[7] A. P. Read, T. Strachan ," Human molecular genetics 2," New York: Wiley;ISBN 0-471-33061-2. Chapter 18: Cancer Genetics, 1999

[8] M. Hollstein, D. Sidransky, B. Vogelstein, C. C. Harris, "p53 mutations in human cancers," Science , 253 (5015): 49–53. doi:10.1126/science.1905840.PMID 1905840, 1991.

[9] C. A. Schmitt, J. S. Fridman, M. Yang, E. Baranov, R. M. Hoffman, S. W. Lowe SW, "Dissecting p53 tumor suppressor functions in vivo," Cancer Cell 1 (3): 289–298. doi:10.1016/S1535-6108(02)00047-8. 2002.

[10] B. Vogelstein, D. Lane, A. J. Levine, "Surfing the p53 network," Nature, vol. 408, pp. 307-310, 2000.

[11] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, "STRING v10: protein-protein interaction networks, integrated over the tree of life," Nucleic Acids Res. 2015 Jan;43 (Database issue):D447-52. doi: 10.1093/nar/gku1003, 2015.

[12] B. Liaoa, K. Dingb "A 3D graphical representation of DNA sequences and its application," Theoretical Computer Science, vol. 358, pp. 56 – 64. doi:10.1016/j.tcs.2005.12.012, 2012.

[13] F. I. Bai, Y. Z. Liu, T. M. Wang TM, "A representation of DNA primary sequences by random walk," Mathematical Biosciences 209, Issue 1, 282–291. doi:10.1016/j.mbs.2006.06.004, 2007.

[14] J. K. Das, P. Pal Choudhury, "Chemical property based sequence characterization of PpcA and its homolog proteins PpcB-E: A mathematical approach," PLoSONE 12(3): e0175031,2016.

[15] J. K. Das, P. Das, K. K. Ray, P. Pal Choudhury, S. S. Jana, "Mathematical Characterization of Protein Sequences Using Patterns as Chemical Group Combinations of Amino Acids," PLoSONE. 8;11(12):e0167651, 2016.

[16] P. Basak, S. Maitra-Majee, J. K. Das, A. Mukherjee, S. G. Dastidar P. Pal Choudhury, A. L. Majumder, "An evolutionary analysis identifies a conserved pentapeptide stretch containing the two essential lysine residues for rice L-myo-inositol 1-phosphate synthase catalytic activity," PLoSONE 12(9):e0185351, 2017.