



Hub Feature Extraction of Tumor Proteins Using Similarity Measurements

Sajeev JResearch Scholar,
Manonmaniam Sundaranar University
Thirunelveli, Tamilnadu, India**Dr. T. Mahalakshmi**Professor and Principal,
Sree Narayana Institute of Technology
Vadakkavila, Kollam, Kerala, India

Abstract— This paper presents a method to extract the hub features of tumor proteins using Similarity Measurements such as Jaccard, Cosine, Dice and Overlap. Through this paper we are not trying to predict the class of a protein as hub proteins or tumor proteins but to prove that a good majority of tumor proteins are hubs. Hydrophobicity, one of the important physio-chemical characteristics of the amino acid along with 28 other physio-chemical characteristics are used for this purpose. Two databases such as Human Protein Reference Database (HPRD) and P53 interaction database are used for testing the method. The application of the proposed method on the random samples of both databases has revealed 72% correctness.

Keywords-Hub protein, TP53, Tumor Proteins, Jaccard, Cosine, Dice, Overlap, Hydrophobic, PIN, Degree of Connectivity, HPRD

I. INTRODUCTION

Among the different types of proteins, Hub proteins are that class of proteins having high degree of interaction in its interaction network. They participate in significant number of protein interactions and play a very important role in the organization of cellular protein interaction pathways [1][2][3].

Among the tumor proteins P53 plays a very important role. P53 (also known as tumor protein 53) is a protein in humans encoded by the TP53 gene which is a tumor suppressor gene, i.e., its activity stops the formation of tumors [4]. Over the years P53 has been shown to interact with more than hundred proteins, which is evident from the pathway information [5].

Most of the biological pathways and processes are believed to be directed by complex protein interactions and are controlled by Hubs [6]. If these proteins are disrupted it can lead to biological lethality [4]. So understanding the characteristic of Hub proteins may in turn increase the significance in understanding the causes of diseases.

Even though lots of studies have been carried out in the fields of Tumor Proteins and Hub Proteins, Hub characteristics of tumor proteins have not attracted much attention so far.

Biological sequence processing has utilized a lot of statistical methods with varying percentage of accuracies in the classification of hub proteins, prediction of protein interaction etc [7][8][9]. The proposed method uses an important statistical technique known as similarity measurements such as Jaccard, Cosine, Dice and Overlap. They are the common similarity measurements which are used in the proposed method to characterize Hub feature in Tumor proteins[10]. In the proposed method feature vectors are generated for both tumor and hub proteins consisting of 29 physio-chemical characteristics and similarity measurements are used to test the similarity between the feature vectors employing various combinations.

Two types of data sets are used in this paper for testing the proposed method. One is from the Human Protein Reference Database (HPRD)[11]. Random samples of human proteins were selected from this database to test the proposed method. The second data set is the P53 interaction database which is obtained from literature [12]. This data set is named as P53 Interaction Data for convenience.

The application of the proposed method on the random samples of both databases has revealed 72% correctness. This method will certainly prove to be useful in Cancer Research involving protein sequence information only.

II. BACKGROUND

This session describes briefly the biological background pertinent to this paper.

A. Hub Protein

Protein interactions are generally represented in the form of a network known as Protein Interaction Network (PIN)[13]. These are visualized in the form of graphs with nodes representing proteins and edges representing interactions between them. One of the attributes associated with each protein in PIN is the connectivity measure. This is the count of number of interactions that a protein has with other members (proteins) of the network. An example of a PIN is given in figure 1.

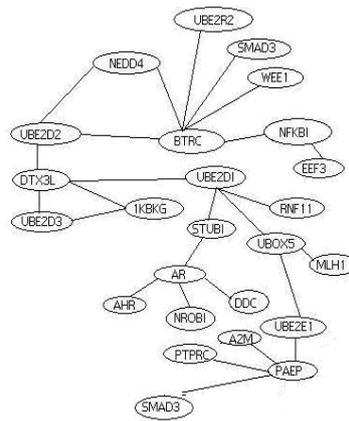


Fig. 1 A sub-network of PIN from HPRD

The protein set used in the PIN given in figure 1 is taken from the HPRD[11]. Only a very small subset of proteins from this database is used for the construction of the PIN. In figure 1 even though the protein identified by BTRC has a connectivity measure of 18, it has been limited to 6 due to space constraints. The protein identified by DDC is a terminal node since its connectivity measure is one. In table I the actual degree of connectivity of the set of proteins used in figure1 is given.

TABLE I DEGREE OF CONNECTIVITY OF PROTEINS IN THE SUBNETWORK GIVEN IN FIGURE 1.

ID	frequency	ID	frequency	ID	frequency
STUB1	29	BTRC	18	AHR	27
UBE2R2	3	1KBKG	57	NR0B1	11
PAEP	9	RNF11	61	DDC	1
DTX3L	7	NFKB1	72	A2M	27
UBE2D3	12	UBOX5	7	PTPRC	44
UBE2D1	9	UBE2E1	12	VBE2G1	6
SMAD3	184	IMMT	37	AR	138
WEF1	20	MCH1	6	UBE2D2	6

It can be seen from the sub network given in figure 1 and from the information given in table I that an interaction between two proteins can be either direct or indirect. In a direct interaction two proteins are connected with an edge. In the case of an indirect interaction two proteins are connected through one or more proteins. As an example in figure 1 UBE2D1 directly interacts with STUB1 where as AR has an indirect interaction with UBE2D1 through STUB1

PIN belongs to the category of Scale Free Networks which has the property that only a few proteins have high connectivity measure [6][13]. This is because the connectivity measure in a Scale Free Network follows a power law distribution [7]. There is an ongoing debate on the threshold value of connectivity measure which is mainly used to classify a protein as Hub or non-Hub[14]. Different methods in this area propose the threshold of connectivity based on various techniques. In the proposed method the statistical property of the connectivity measure is used to find the threshold value and used to characterize Hub proteins.

B. Similarity Measurements

The similarity between two tuples t_i and t_j , $sim(t_i, t_j)$, in a database D is a mapping from $D \times D$ to the range [0, 1]. Thus, $sim(t_i, t_j) \in [0, 1]$.

The objective is to define the similarity mapping such that documents that is more alike have a higher similarity value [15]. Thus, the following are desirable characteristics of a good similarity measure.

$$\forall t_i \in D, sim(t_i, t_i) = 1$$

$$\forall t_i, t_j \in D, sim(t_i, t_j) = 0 \text{ if } t_i \text{ and } t_j \text{ are not alike at all}$$

$$\forall t_i, t_j, t_k \in D, sim(t_i, t_j) < sim(t_i, t_k) \text{ if } t_i \text{ is more like } t_k \text{ than it is like } t_j \text{ [15]}$$

Here four types of similarity measures are used for finding similarity of sequences. They are:

Dice measure

$$sim(t_i, t_j) = \frac{2 \sum_{h=1}^k t_{ih} t_{jh}}{\sum_{h=1}^k t_{ih}^2 + \sum_{h=1}^k t_{jh}^2} m$$

Jaccard measure

$$\text{sim}(t_i, t_j) = \frac{\sum_{h=1}^k t_{ih} t_{jh}}{\sum_{h=1}^k t_{ih}^2 + \sum_{h=1}^k t_{jh}^2 - \sum_{h=1}^k t_{ih} t_{jh}}$$

Cosine measure

$$\text{sim}(t_i, t_j) = \frac{\sum_{h=1}^k t_{ih} t_{jh}}{\sqrt{\sum_{h=1}^k t_{ih}^2} \sqrt{\sum_{h=1}^k t_{jh}^2}}$$

Overlap measure

$$\text{sim}(t_i, t_j) = \frac{\sum_{h=1}^k t_{ih} t_{jh}}{\min(\sum_{h=1}^k t_{ih}^2, \sum_{h=1}^k t_{jh}^2)}$$

In these formulas it is assumed that similarity is being evaluated between two vectors $t_i = (t_{i1}, t_{i2}, \dots, t_{ik})$ and $t_j = (t_{j1}, t_{j2}, \dots, t_{jk})$ and vector entries usually are assumed to be nonnegative numeric values [15].

C. Tumor Protein – TP53

Tumor protein P53, also known as P53 or tumor suppressor P53, is a protein that in humans which is encoded by the gene TP53. As a tumor suppressor, the TP53 protein is crucial in regulating the cell cycle and, thus, preventing cancer. As such, P53 has been described as "the guardian of the genome" because of its role in preserving stability by avoiding genome mutation. Hence TP53 is also classified as a tumor suppressor gene [16][17][18][19].

p53 is also known as cellular tumor antigen p53 (UniProt name), antigen NY-CO-13, phosphoprotein p53 and transformation-related protein 53 (TRP53). In humans, the TP53 gene is located on the short arm of chromosome 17 (17p13.1) [16][17][18][19].

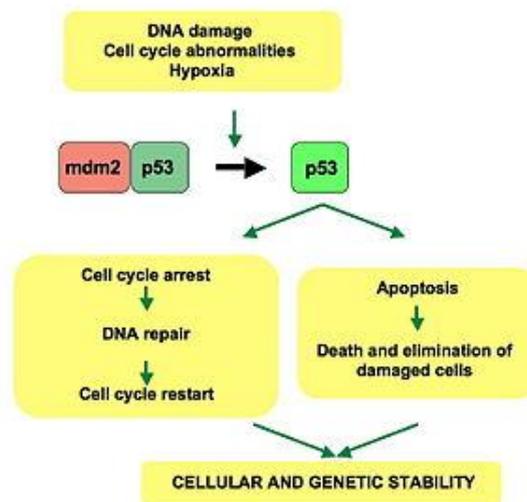


Fig. 2 TP53 Protecting cell from DNA damage

If the TP53 gene is damaged, tumor suppression will be severely compromised. The TP53 gene can also be modified by mutagens (chemicals, radiation, or viruses), increasing the likelihood for uncontrolled cell division. More than 50 percent of human tumors contain a mutation or deletion of the TP53 gene [20]. Pharmacological reactivation of p53 presents itself as a viable cancer treatment option [21][22]. Loss of P53 creates genomic instability that most often results in an aneuploidy phenotype [23].

TP53 is inactivated by its negative regulator, mdm2 in a normal human cell. Due to DNA damage various pathways will lead to the dissociation of the p53 and mdm2 complex. Upon activation, p53 will induce a cell cycle arrest to allow either repair or survival of the cell or apoptosis to discard the damaged cell [24]. This concept is depicted in the figure 2.

III. DATA SET

In this paper two sets of data have been used to test the proposed method. The first set is taken from HPRD [11]. From the database whole set of human protein ID's were obtained. In this database there were 27080 human proteins. Among them 9630 have interactions with others. This information is presented in the form of binary interactions in the database. From this, it was possible to find the count of number of interactions of each protein. This count is taken as the degree of connectivity of the protein and it ranged from 0 to 267. The table below gives the number of proteins having the degree of connectivity k.

TABLE III DEGREE OF CONNECTIVITY VS PROTEIN FREQUENCY IN HPRD

Degree of Connectivity (k)	Number of proteins
0	17450
1	2237

2	1424
3	1009
4	759
5	618
6	468
7	422
8	287
>> 8	2406
Total	27080

It can be seen from the table III that as the value of k increases the frequency of the protein decreases. Using this information as a frequency table, its mean was calculated and was obtained as 8.0557. It is again evident from the table III that frequency count of the proteins with $k < 9$ is 7224. That is there are 2406 proteins with $k > 8$ which is around 25% of the total interacting proteins. In the proposed method the threshold for connectivity of hub proteins for this database is taken as 8 based on the above analysis.

The second set is taken from the interacting protein set of P53 protein[12]. In this database there were 108 proteins which have shown interaction with P53. The sequence information is obtained from the NCBI Database [25].

IV. PROPOSED METHOD

Proposed method gives a better and more robust similarity measure to check how far one sequence is similar to another one in terms of the physicochemical properties of the sequences. The proposed method is divided into a two steps which are listed below.

Step 1) Feature vector creation

Feature vectors are used to characterize objects in real world [26]. These feature vectors undergo data mining which in turn help researchers to come out with accurate results. Feature vectors are essential in multi-dimensional analysis. Feature reduction is necessary if the dimensions go beyond a particular limit which increases the computational complexity. It is always a challenge to derive simple but useful features out of objects under consideration. One such object is protein sequence. Hence our main focus in the proposed method is protein sequence.

In Step 1, the selected human proteins from HPRD and P53 Interaction Database are used to derive feature vectors.

Two types of feature vectors are generated from protein sequences. The components of the first feature vector are the counts of 20 different amino acids present in the sequence [27]. The same feature vectors were successfully applied by Eun-Joon Par et. al in the prediction of protein sub cellular locations with the help of support vector machines [28]. In another work on protein-protein interaction prediction, amino acid feature vectors have given highly accurate results [29].

Second feature vector contains nine components. These components are the counts of Hydrophobic, Aromatic, Positive, Negative, Charged, Polar, Tiny, Small and Aliphatic amino acids in the protein sequence [30]. These components are the important physicochemical properties of the amino acids. The components fall under three classes such as polarity, hydrophobicity and size. Venn diagram in Figure 3 depicts these classes. These features were successfully used by K.S.M. Tozammel Hossain et.al to induce Graphical Models of Residue Couplings[30].

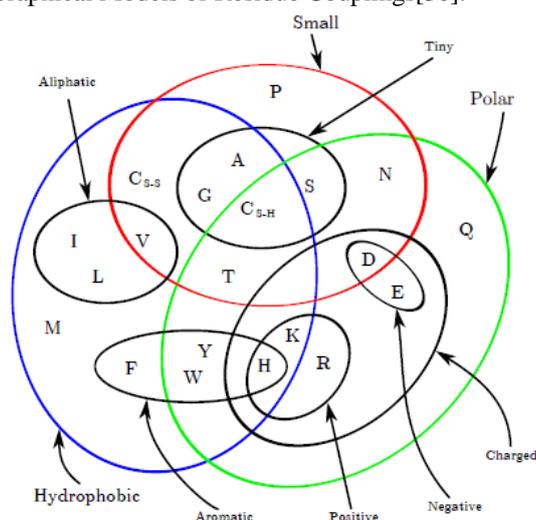


Fig. 3 Different classes of proteins

Nine components along with the three classes are listed in the table III. In the table first column gives the class name, second column gives the feature and third column explains the different amino acid residues which fall under a particular feature. One amino acid residue can be part of more than one features. For example amino acid 'E' belongs to features Polar, Negative and Charged. Another amino acid K belongs to features Polar, Positive, Charged and Hydrophobic.

TABLE III AMINO ACID RESIDUES ALONG WITH THEIR CORRESPONDING CLASSES AND FEATURES

Class	Feature	Amino Acid Residues
Polarity	Polar	Q, D, E, H, K, R, Y, W, T, C, S, N
Polarity	Positive	K, H, R
Polarity	Negative	D, E
Polarity	Charged	D, E, K, H, R
Size	Small	P, A, C, A, G, S, N, D, T, V
Size	Tiny	A, G, S, C
Hydrophobicity	Aliphatic	I, L, V
Hydrophobicity	Aromatic	F, W, Y, H
Hydrophobicity	Hydrophobic	A, C, G, T, H, K, Y, W, F, M, I, L, V

Step 2) Application of feature vectors on similarity measurements

In step 2 the feature vectors generated from the protein sequences are tested with all the four similarity measures - Dice, Jaccard, Cosine and Overlap.

For the better understanding, the proposed method is illustrated with four dummy protein sequences in which two of them are similar and the rest are dissimilar.

Table IV given below contains the sets of similar and dissimilar dummy protein sequences.

TABLE IV SIMILAR AND DISSIMILAR DUMMY PROTEIN SEQUENCES

Similar protein sequences	
Name	Sequence
Seq1	ACCPKMDLTIMCTETFFHHGLACMKKNNMPDDQCQDDTSSRRD DVWVYKKPWYLLMPQSDWYALMCTEFGHG
Seq2	TPCCLMMNQSRWYYWMNDACCPTTLDVEFKLGHH AAEFELMMWVSTGHHIACDDPSPDBRTST
Dissimilar protein sequences	
Name	Sequence
Seq3	ACCPKMDLTIMCTETFFHHGLACMKKNNMPDDQCQDDTS SRRDDVWVYKKPWYLLMPQSDWYALMCTEFGYHG
Seq4	PALSMNCFAIMQENRNESRWMYERMNISINDFFEARML VEFNSKLMALKRLFAAENFEFVLLMSMVSTGHIAID

From the sequences given in the table IV we derive the amino acid counts. Table V and table VI contains the amino acid counts for similar and dissimilar dummy protein sequences.

TABLE V AMINO ACID COUNTS FROM THE SIMILAR DUMMY PROTEIN SEQUENCES.

Amino Acid counts for the residues	
	A C D E F G H I K L M N P Q R S T V W Y
Seq1	3 6 9 2 3 3 3 1 5 5 6 2 4 4 2 3 5 1 5 3
Seq2	4 5 5 3 3 2 4 1 1 5 5 2 4 2 2 4 6 2 3 2

TABLE VI AMINO ACID COUNTS FROM THE DISSIMILAR DUMMY SEQUENCES.

Amino Acid counts for the residues	
	A C D E F G H I K L M N P Q R S T V W Y
Seq1	3 6 9 2 3 3 3 1 5 5 6 2 4 4 2 3 5 1 5 4
Seq2	7 1 2 7 7 1 1 5 2 7 8 7 1 1 5 6 1 3 1 1

The counts obtained in the previous tables are applied on the similarity measures for similarity checking. Table VII and table VIII given below shows the results obtained after that process

TABLE VII RESULTS OF PROPOSED METHOD ON SIMILAR DUMMY SEQUENCES.

Measure	Value Obtained	% of similarity
Dice	0.917491749	91.7%
Jaccard	0.847560976	84.8%

Cosine	0.930244654	93%
Overlap	1.098814229	100%

TABLE VIII RESULTS OF PROPOSED METHOD ON DISSIMILAR DUMMY SEQUENCES.

Measure	Value Obtained	% of similarity
Dice	0.6435897	64.4%
Jaccard	0.4744802	47.4%
Cosine	0.6455023	64.6%
Overlap	0.6972222	69.7

The application of real data sets consisting of feature vectors with 29 components is elaborated in the next section.

V. RESULTS AND DISCUSSION

Proposed method for analyzing the similarity between human proteins with the help of similarity measurements has proved good results using sample sets of proteins selected randomly from the HPRD and TP53 interaction data.

For analysis five sample sets of data are selected. Each sample set contains two sets and each contains 18 protein sequences.

There are many measures used to test the correctness of algorithms. Sensitivity, Specificity, Accuracy, Correctly Classified Instance (CCI) are some of such measures. Expressions for such measure are given below.

Sensitivity	$TP / (TP + FN)$
Specificity	$TN / (TN + FP)$
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
CCI	$(TP + TN) / N$

Here TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives and N is the total number of instances.

In our proposed method CCI is used. CCI is a measure which is used to find the number of correctly classified instances to the total number of instances [31].

Sample sets containing 18 random samples of interacting data belonging to HPRD data set were compared with another 18 samples of the interacting data set, they have shown positive results. Five such sample sets were tested which show a correctness of 71.6%. This is illustrated by table IX below.

TABLE IX RESULTS OF PROPOSED METHOD ON HPRD INTERACTION DATA SETS.

Set1 - Interaction data of HPRD -18					
Set2 - Interaction data of HPRD -18					
Sample Sl. No	Dice	Jaccard	Cosine	Overlap	CCI
1	73	72	73	76	74%
2	75	74	76	77	76%
3	67	67	68	68	68%
4	63	63	65	66	64%
5	75	74	76	77	76%

According to this result in Table IX when five sample sets were applied on the similarity measure we got an average of 71.6% result. The minimum value obtained is 63% and maximum value obtained is 77%.

When the same number of random samples consisting of non interacting proteins were applied on this method they too have shown positive results. Five such sample sets were tested which show a correctness of 67.2%. This is illustrated by the table X.

As per the result in Table X when five sample sets were applied on the similarity measure we got an average of 71.6% result. The minimum value obtained is 58% and maximum value obtained is 76%.

TABLE X RESULTS OF PROPOSED METHOD ON HPRD NON-INTERACTION DATA SETS.

Set1 - Non interaction data of HPRD -18					
Set2 - Non interaction data of HPRD -18					
Sample Sl. No	Dice	Jaccard	Cosine	Overlap	CCI
1	62	60	63	64	62%
2	69	68	70	72	70%
3	60	58	62	63	61%
4	70	69	71	71	70%
5	72	70	75	76	73%

The same process was repeated with 18 samples of P53 interaction data and 18 samples of HPRD interacting data set which has shown positive results. Five such sample sets were tested which show a correctness of which accounts to 72.6%. This is illustrated in the table XI.

TABLE XI RESULTS OF PROPOSED METHOD ON HPRD INTERACTION DATA AND TP53 DATA SETS.

Set1 - Interaction data of HPRD -18					
Set2 - Interaction data of TP53 -18					
Sample Sl. No	Dice	Jaccard	Cosine	Overlap	CCI
1	71	69	73	75	72%
2	73	72	74	75	74%
3	67	66	67	69	67%
4	76	74	76	77	76%
5	73	73	74	76	74%

From the Table XI when five sample sets were applied on the similarity measure we got an average of 72.6% result. The minimum value obtained is 66% and maximum value obtained is 77%.

When the same number of random samples consisting of non interacting proteins of HPRD and TP53 interaction data were applied on this method only 48.5% correctness was shown which indicates the dissimilarity between interacting and non interacting protein sequences. This is illustrated in the table XII.

TABLE XII RESULTS OF PROPOSED METHOD ON HPRD NONINTERACTION DATA AND TP53 DATA SETS.

Set1 - Non interaction data of HPRD -18					
Set2 - Interaction data of TP53 -18					
Sample Sl. No	Dice	Jaccard	Cosine	Overlap	CCI
1	42	42	43	46	43%
2	43	42	44	46	44%
3	51	50	52	53	51%
4	52	51	54	55	53%
5	37	36	39	41	38%

According to the above result in Table XII when five sample sets were applied on the similarity measure we got an average of 48.5% result. The minimum value obtained is 36% and maximum value obtained is 55%.

So it is evident from these findings that TP53 interaction data has more similarities with interacting proteins of HPRD and showing more hub features. At the same time Non interaction proteins have shown less interaction with TP53 interaction data.

VI. CONCLUSION

Protein interactions are ubiquitous and essential for cellular functions. Even though tumor proteins were considered more vital compared with other proteins, no such studies were carried out to fish out its key role in PIN networks.

The proposed method is a two stage process which uses physic-chemical properties of proteins. The application of this method on random sets of human proteins of HPRD database and TP53 interaction data has shown satisfactory results.

As a continuation of this work, we think the proposed method can be applied to study the samples of Passenger and Driver Proteins which are vital in cancer research [32]. Hope such studies will directly or indirectly help to further cancer research experiments harnessing the opportunities of Proteomics.

REFERENCES

- [1] Michael Hsing, Kendall Grant Byler and Artem Cherkasov, "The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks", BMC Systems Biology, 2008, 2:80
- [2] Barabasi A L and Oltvai Z N, "Network biology: understanding the cell's functional organization", Nat Rev Genet 2004, 5(2):101-113.
- [3] Sriganesh Srihari et al, "Detecting Hubs and Quasi Cliques in Scale-free Networks", IEEE, 2008.
- [4] Alexei Vazquez, Elisabeth E Bond, Arnold J Levine, Gareth L Bond, "The genetics of the p53 pathway, apoptosis and cancer therapy", Nature Reviews Drug Discovery (2008) Volume: 7, Issue: 12, Publisher: Nature Publishing Group Pages: 979-987.
- [5] Carol Prives, 2. Peter A. Hall, "The p53 pathway", The Journal of Pathology, Special Issue: Molecular and Cellular Themes in Cancer Research", Volume 187, Issue 1, pages 112-126, January 1999.
- [6] Albert R, "Scale-free networks in cell biology", J Cell Sci 2005, 118 (Pt 21) : 4947-4957.
- [7] Nizar N. Bataba, Laurence D. Hurst and Mike Tyers, "Evolutionary and Physiological Importance of Hub Proteins", PLOS Computational Biology, 2006, 2, 7, 748:756

- [8] Sumeet Agarwal et. al., “Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks”, PLOS Computational Biology, June 2010, Volume 6, Issue 6.
- [9] Ramon Aragues et. al., “Characterization of Protein Hubs by Inferring Interacting Motifs from Protein Interactions”, PLOS Computational Biology, September 2007, Volume 3, Issue 9.
- [10] Vikas Thada et.al, “Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm”, International Journal of Innovations in Engineering and Technology (IJET), SSN:2319-1058.
- [11] <http://www.hprd.org/> Release 9 dated May 24, 2010
- [12] <http://en.wikipedia.org/wiki/P53> dated 18/9/2011 at 9.00 a.m.
- [13] S. Wutchy, “Scale-free behavior in protein domain networks”, Mol. Bio. Evolution 18, 2001.
- [14] Michael Hsing, Kendall Grant Byler and Artem Cherkasov, “The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks”, BMC Systems Biology, 2008, 2:80
- [15] Margaret H. Dunham et. al, “Data Mining – Introductory and Advanced Topics”, ISB 978-81-7758-785-2
- [16] Matlashewski G, Lamb P, Pim D, Peacock J, Crawford L, Benchimol S. “Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene”. EMBO J.. 1984;3(13):3257–62. PMID 6396087.
- [17] Isobe M, Emanuel BS, Givol D, Oren M, Croce CM. Localization of gene for human p53 tumour antigen to band 17p13. Nature. 1986;320(6057):84–5. doi:10.1038/320084a0. PMID 3456488.
- [18] Kern SE, Kinzler KW, Bruskin A, Jarosz D, Friedman P, Prives C, Vogelstein B. Identification of p53 as a sequence-specific DNA-binding protein. Science. 1991;252(5013):1708–11. doi:10.1126/science.2047879. PMID 2047879.
- [19] McBride OW, Merry D, Givol D. The gene for human p53 cellular tumor antigen is located on chromosome 17 short arm (17p13). Proc. Natl. Acad. Sci. U.S.A.. 1986;83(1):130–134. doi:10.1073/pnas.83.1.130. PMID 3001719.
- [20] Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science. 1991;253(5015):49–53. doi:10.1126/science.1905840. PMID 1905840.
- [21] Ventura, Andrea; David Kirsch; Margaret McLaughlin; David A. Tuveson; Jan Grimm; Laura Lintault; Jamie Newman; Elizabeth E. Reczek; Ralph Weissleder; Tyler Jacks (8 February 2007). "Restoration of p53 function leads to tumour regression in vivo". International weekly journal of science 445 (7128): 661–665. doi:10.1038/nature05541. PMID 17251932. Retrieved 2013-03-29.
- [22] Patricia A. J. Muller & Karen H. Vousden, "p53 mutations in cancer", Nature Cell Biology, 15, 2–8 (2013) doi:10.1038/ncb2641
- [23] Herce, HD; Deng, W; Helma, J; Leonhardt, H; Cardoso, MC (2013). "Visualization and targeted disruption of protein interactions in living cells.". Nature communications 4: 2660. PMID 24154492.
- [24] Hock AK, Vigneron AM, Carter S, Ludwig RL, Vousden KH (2011). "Regulation of p53 stability and function by the deubiquitinating enzyme USP42". EMBO J 30 (24): 4921–30. doi:10.1038/emboj.2011.419.
- [25] <http://www.ncbi.nlm.nih.gov/> dated 18/9/2014.
- [26] Xing-Ming Zhao et. al, “A novel approach to extracting features from motif content and protein composition for protein sequence classification”, Neural Networks (2005) 1019-1028, Elsevier Ltd.
- [27] <http://www.sigmaaldrich.com/life-science/metabolomics/learning-centre/amino-acid-reference-chart.html> dated Oct 18 2014.
- [28] Eun-Joon Par et. al., “Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs”, Bioinformatics, Volume 19, Issue 13, Pp. 1656-1663
- [29] Joel R. Bock et.al., “Predicting protein-protein interactions from primary structure”, Bioinformatics, Volume 17, Issue 5, Pp 455
- [30] K.S.M. Tozammel Hossain et.al, “Using Physicochemical properties of Amino Acids to induce Graphical Models of Residu Couplings”, BIOKDD 2011 Aug 2011 San Diego, CA USA ACM 978-1-4503-0839-7.
- [31] Vidya A et. al, “CFS with Combined Search Methods for Dimensionality Reduction in Classifying Aging and Not-aging related DNA Repair Genes”, International Journal of Computational Intelligence Research, ISSN 0973-1873, Vol 10
- [32] Tao Meng et. al, “Wavelet Analysis in Current Cancer Genome Research : A Survey”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 10, No 6, November/December 2013 .