ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Identification of Similar Ligands Using Microarray Data for BER Genes.

Udayakumar Mani*, and Sai Mukund Ramakrishnan.

Department of Bioinformatics, SASTRA University, Tanjore – 613 401, Tamil Nadu, India.

ABSTRACT

Ligands are small molecules which act as a switch to turn off or on the function of the protein. A misfunction of such ligand may affect the function of the Gene which could be altered using similar ligands. The domain for the similar ligands can be identified using Microarray, which is the high throughput technique used for measuring expression levels of very short sections of a Gene. Misfunction of DNA repair genes causes major diseases, one such mechanism that is taken under consideration is Base Excision Repair (BER). In order to find the similar ligands for DNA repair Genes, we designed a tool where the gene name is inputted. A cluster plot is generated for the protein products of the Gene inputted with respect to its functions. Using the microarray data a high expression profile is constructed and the domain is extracted. The list of possible similar ligands is obtained along with the domain, which forms their site of action. We have thus stratified a methodology to accomplish all these tasks. Results had 95% and above similarity with original ligands. It thereby reduces the search time to a greater extent. The plots comfort the user with better understanding about the Gene products and its expressions levels.

Keywords: DNA Repair Mechanism, BER Genes, Clustering Plot, Ligand Identification, Affimetrix data.

**Corresponding author*

INTRODUCTION

DNA in the living cell is subject to many chemical alterations. If the genetic information encoded in the DNA is to remain uncorrupted these alterations must be corrected. There are 4 different kinds of DNA damage. They are: All four of the bases in DNA (A, T, C and G) can be covalently modified at various positions example: C being converted to a U; Mismatches of the normal bases because of a failure in proofreading during DNA replication example: incorporation of the pyrimidine U (normally found only in RNA) instead of T; Breaks in the backbone can be limited to one of the two strands or on both strands; Crosslink covalent linkages can be formed as an interstrand or an intrastrand on the same DNA [1]. These repair genes pave way for 2 types of repair mechanisms: Direct chemical reversal of the damage and Excision Repair mechanisms. There are three modes of excision repair, each of which employs specialized sets of enzymes: Base Excision Repair (BER); Nucleotide Excision Repair (NER); Mismatch Repair (MMR). Out of these three excision repair, we mainly focus on BER genes in order to prevent the cause of major diseases.

Spontaneous hydrolytic de-purination and deamination of cytosine and 5-methylcytosine residues, multiple reactions with hydroxyl free-radicals generated as accidental by-products of normal oxygen metabolism and formation of covalent DNA adducts on exposure to reactive small metabolites and coenzymes generate a variety of DNA lesions that require precise and rapid repair. The main strategy for correcting such DNA damage is base excision repair (BER). An altered DNA base is excised in its free form by a DNA glycosylase and the resulting abasic site is corrected by the concerted action of an AP endonuclease, a DNA polymerase and a DNA ligase [2].

BER is important for removing damaged bases that could cause mutations by either impairing or can lead to breaks in the DNA during replication. BER is initiated by DNA glycosylases, which recognize and remove specific damaged or inappropriate bases, forming apurinic or apyrimidic or abasic sites. These are then cleaved by an AP endonuclease. The resulting single-strand break can then be processed by either short-patch (where a single nucleotide is replaced) or long-patch BER (where 2-10 new nucleotides are synthesized) [3].

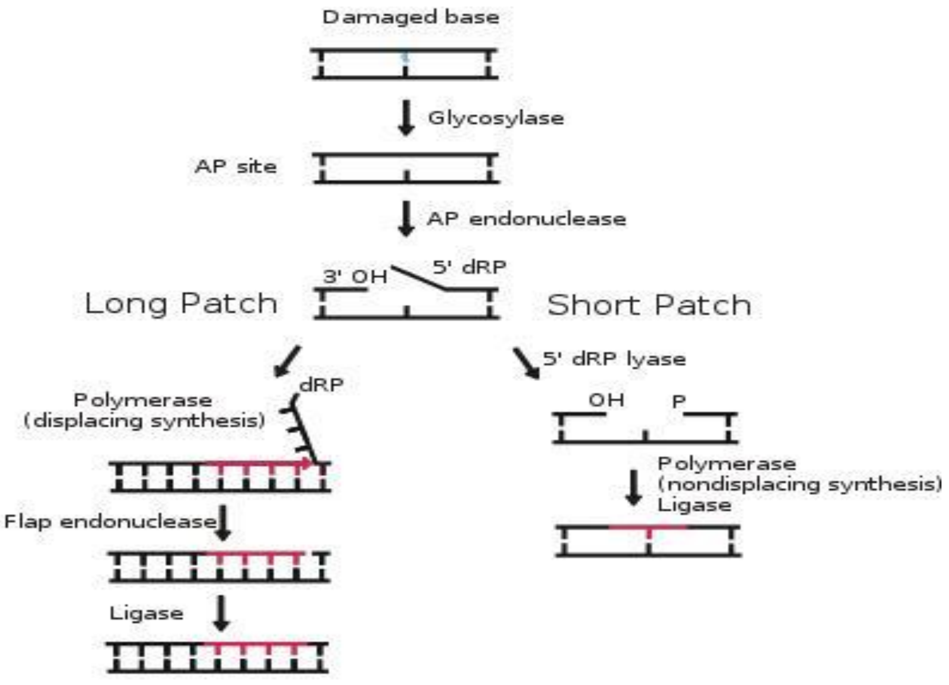


Fig. 1: Base Excision Repair mechanism

The main focus of this paper is to develop a methodology for identifying similar ligands for the ligands which are actually responsible for performing certain roles in dictating a gene's function. Literature has proven that the ligands are responsible for making a switch between the on or off mechanism of the protein's function. The misfunction of ligands responsible for activating these genes may also be a vital cause for the failure of certain biological mechanism **[4]**. Thus the similar ligands may help out to overcome the failure and also to remove the effects to certain extent. Major diseases like accelerating ages and various types of cancer are caused by a failure in the DNA repair mechanism. The BER genes are indeed the most vital ones that get affected. More over the microarray data apart from giving the expression level of the gene also indicates the conserved domain, where the function could most probably take part **[5]**. Thus in devising this algorithm we can solve the above mentioned problems. This methodology is being developed with higher computational performance to depict the results in lesser time with higher accuracy.

METHODOLOGY

Comprehending different tools online and offline, the methodology discussed here has been developed to work in three different combinatorial phases. The user inputs certain known gene ID. Using the prerequisite the back end algorithm identifies all the similar sets of ligands from the NCBI database using its FTP link [6]. Based on clustering algorithms and high expression profiles these similar ligands are subjected to a series of filters out of which only 95% – 98% similar structures are taken as the final output data set. This demonstrates the overview of the entire methodology.

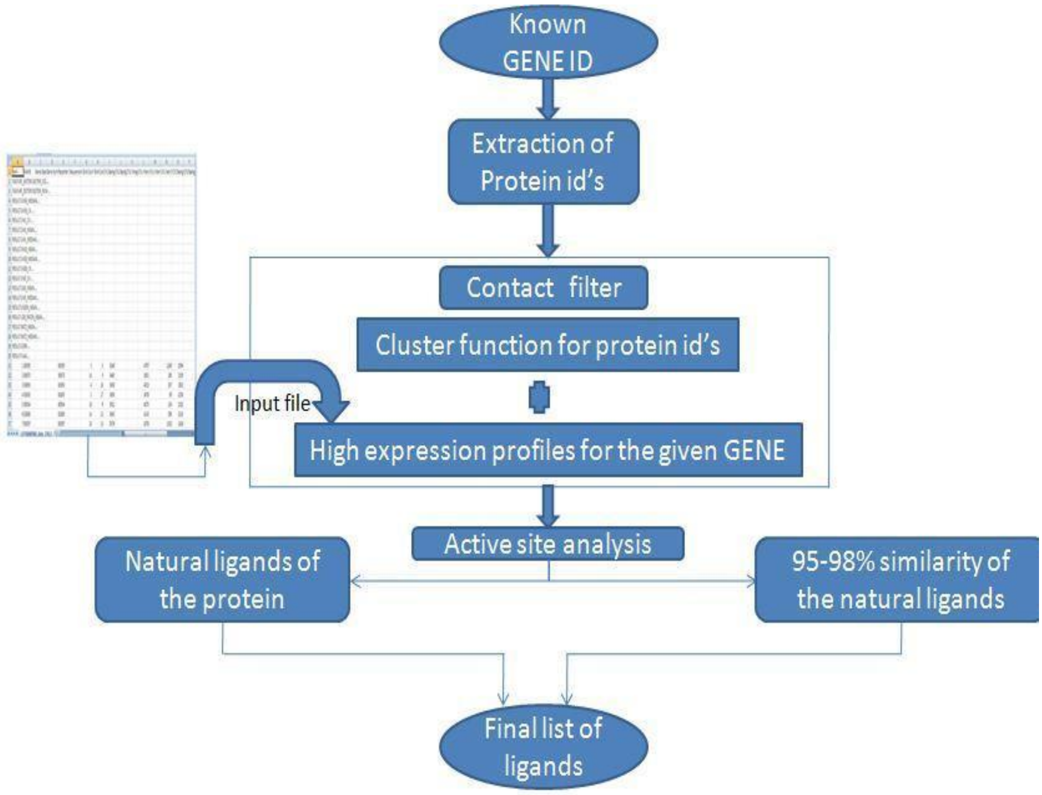

Known GENE ID
Extraction of Protein id's
Contact filter
Cluster function for protein id's
Input file
High expression profiles for the given GENE
Active site analysis
Natural ligands of the protein
95-98% similarity of the natural ligands
Final list of ligands

Fig. 2: Schema of the proposed methodology

The following are the combinatorial phases through which the entire algorithm (methodology) runs to process the output.

Phase I – Cluster Plot:

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression [7]. This data mining approach is made use to cluster the gene products along with its functions and to comprehend the information into a simple graphical plot.

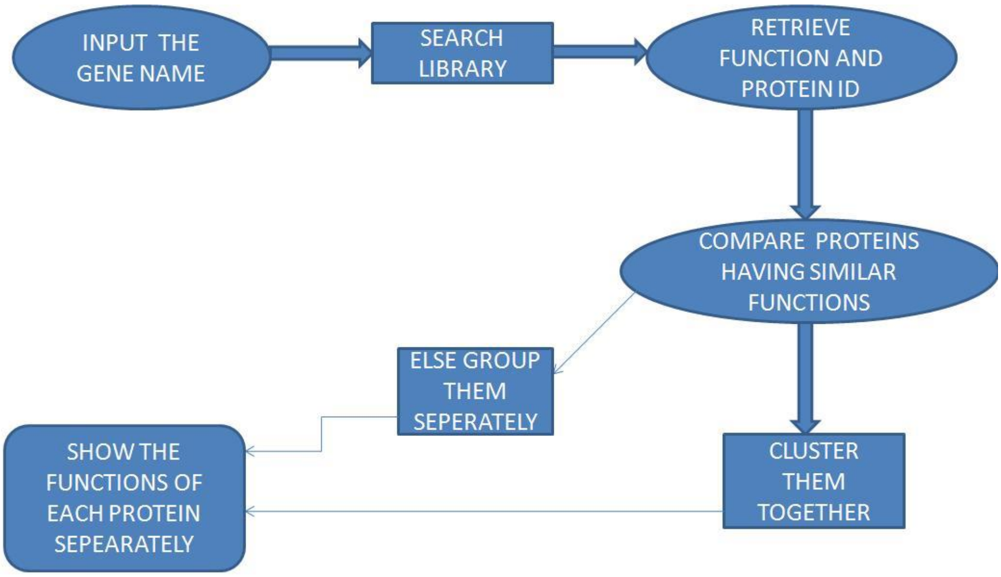

INPUT THE GENE NAME
SEARCH LIBRARY
RETRIEVE FUNCTION AND PROTEIN ID
COMPARE PROTEINS HAVING SIMILAR FUNCTIONS
ELSE GROUP THEM SEPERATELY
CLUSTER THEM TOGETHER
SHOW THE FUNCTIONS OF EACH PROTEIN SEPEARATELY

Fig. 3: Flowchart of the Cluster Plot

Phase II – High Expression Profile:

The microarray experiments are generally carried out to measure the expression level of the gene for a given condition. Thus here microarray data is used to see which potion of the gene has a high expression level. It uses various methods like ANOVA [8] and Tukey's test [9] in order to specifically identify the high expressed probe of the gene which is used in the microarray data. The methodology has been deployed to accept only Affymetrix experimental microarray data [10].

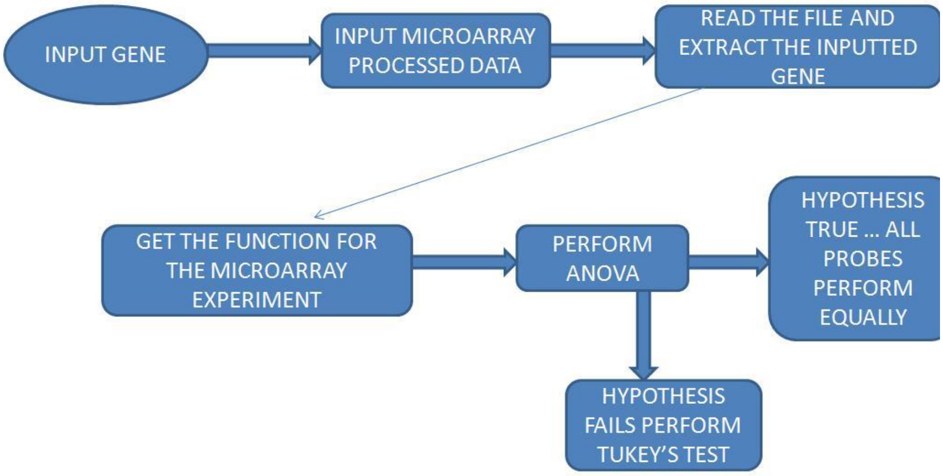

INPUT GENE
INPUT MICROARRAY PROCESSED DATA
READ THE FILE AND EXTRACT THE INPUTTED GENE
GET THE FUNCTION FOR THE MICROARRAY EXPERIMENT
PERFORM ANOVA
HYPOTHESIS TRUE ... ALL PROBES PERFORM EQUALLY
HYPOTHESIS FAILS PERFORM TUKEY'S TEST

Fig. 4: Flowchart of High expression profile

Phase III – Similar Ligand Identification:

For the original set of ligands, the similar ligands were found using tanimoto index and structural similarity greater than or equal to 95%. The tanimoto index is mainly used to identify the similarity between two different items. Thus here tanimoto is used to identify the similarity between the ligand and its similar ligand. Generally the tanimoto index [11] of 0.9 represents the similarity of 95%. After finding the similar ligands they were tested with the help of docking to find whether their site of binding is the same and their interacting residues are the same.

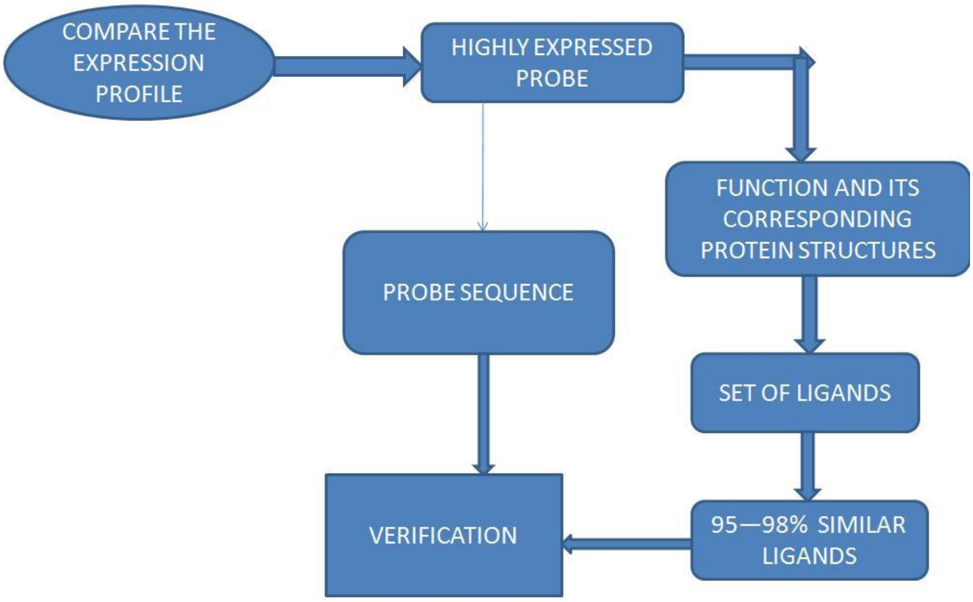
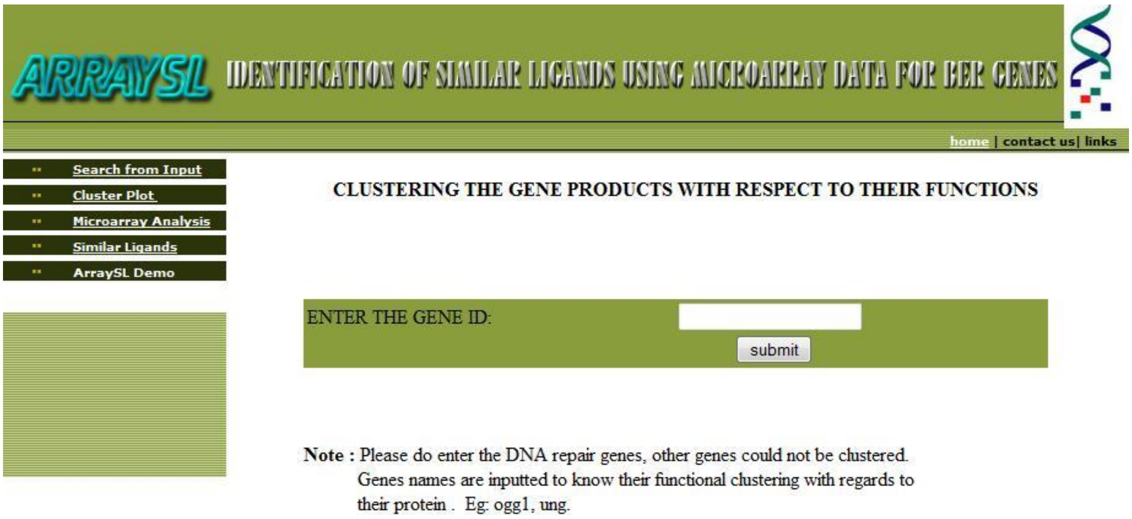


Fig. 5: Flowchart of similar ligand identification

RESULTS

The entire methodology has been divided into 3 phases of modules. They were Cluster plot, high expression profile and the similarity ligand identification respectively.



CLUSTERING THE GENE PRODUCTS WITH RESPECT TO THEIR FUNCTIONS

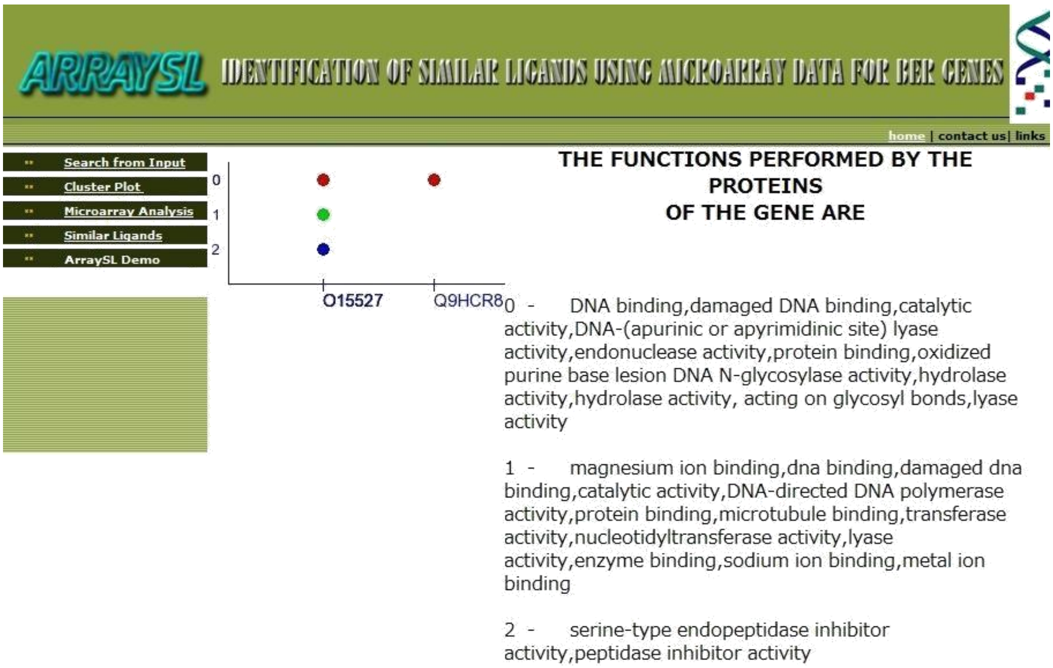
ENTER THE GENE ID:

Note : Please do enter the DNA repair genes, other genes could not be clustered. Genes names are inputted to know their functional clustering with regards to their protein . Eg: ogg1, ung.

Fig. 6: Search form that is used to enter the required inputs

First phase:

In the first phase of the process the cluster plot (Fig: 3) was constructed by analysing the individual protein with respect to their function. Proteins sharing similar functions were grouped together in a straight line which forms a cluster. Each function is indicated in different colours of RGB (red, green, blue) combination. The pictorial representation of the plot is made with the help of GD module [12].



THE FUNCTIONS PERFORMED BY THE PROTEINS OF THE GENE ARE

0 - DNA binding,damaged DNA binding,catalytic activity,DNA-(apurinic or apyrimidinic site) lyase activity,endonuclease activity,protein binding,oxidized purine base lesion DNA N-glycosylase activity,hydrolase activity,hydrolase activity, acting on glycosyl bonds,lyase activity

1 - magnesium ion binding,dna binding,damaged dna binding,catalytic activity,DNA-directed DNA polymerase activity,protein binding,microtubule binding,transferase activity,nucleotidyltransferase activity,lyase activity,enzyme binding,sodium ion binding,metal ion binding

2 - serine-type endopeptidase inhibitor activity,peptidase inhibitor activity

Fig: 7. Cluster plot

Second Phase:

In the second phase of the process the high expression profile (Fig: 4) is to be constructed with the help of the microarray data. In this phase we provide the user with two options, one is the user could directly upload the microarray data from their experiments, but if the data is already available in the microarray database, then the user can use web service scheme that we provide to download the data from the web directly. The expression levels are pictorially represented using GD module. The expression levels of probe set are checked for the particular gene which the user has inputted. The difference among the probe sets and its independency is checked with the help of the ANOVA test and error correction is carried out with the help of Tukey‘s test if the ANOVA hypothesis is proved to be false.

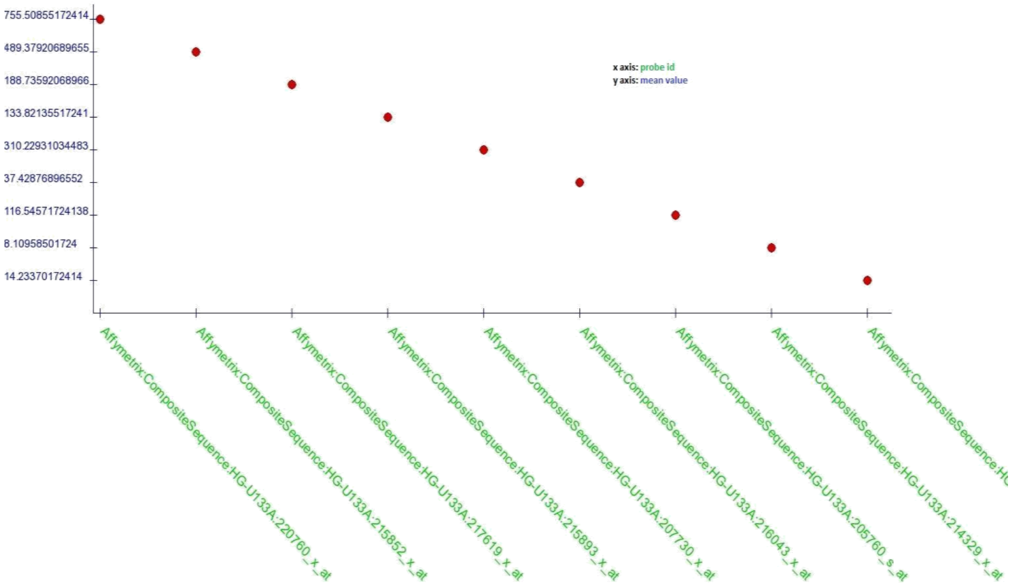


Fig. 8: High expression profile

Third Phase:

In the third phase of the process the similarity ligands gives the following information:

- All the possible PDB codes [13]
- The set of ligands actually present in the protein complex to activate them.
- The key role of the protein, in short the functions of the proteins are listed.
- The probe id which has the high expression is indicated along with its sequential information.

The similarity ligands with respect to the set of original ligands are listed out along with their SMILES notation, molecular formula and the molecular descriptors with respect to RULE OF FIVE (Lipinski rule of 5) [14].

DISCUSSION

The final output has three distinct results. One could observe that the result produced at one phase of the process is mainly helpful for the obtaining the result of the next phase of the process. Thus in short one cannot move forward without the results of phase I.

PHASE I
PHASE II
PHASE III

Fig. 9: Sequential flow of the proposed methodology

The Phase I:

The result obtained in the first phase was a cluster plot. The cluster plot was plotted with the gene name being inputted. With the help of the gene name its gene product (protein) were collected and the function corresponding to each protein was also collected. A comparative analysis was used to find all those proteins which had similar functions. With respect to the function the proteins were plotted and thus obtained a cluster plot. This cluster plot can easily make the user to understand the similarity within proteins with respect to their function. The several functions considered to depict the cluster plot are also listed. (Fig: 6)

The Phase II:

The result obtained in the second phase was a high expression profile. Generally the microarray is a collection of expression values for each sample under a given condition. So as far as the entire data is considered we will have more than one value as they test it under various conditions. Thus in a microarray data one could find a large amount of numerical information. Thus the statistical concept of ANOVA has been used to find out the highly expressed one with the help of Tukey's test. The results were very accurate when they were compared and analyzed using SPSS package [15] (Fig: 7). One important observation in this plot is that the mean values are in y-axis and probe id's in x-axis. The point is that the expression levels of the probes are in no way related to mean values. They only depend on the significance value. In the plot one could observe that the mean values are not in an order. This clearly proves that only the Significance value is being considered.

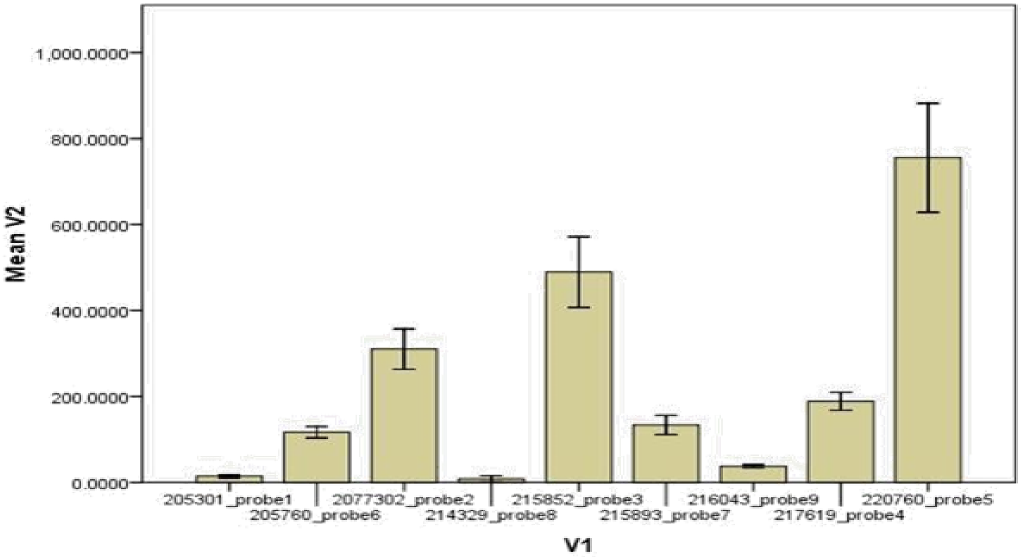
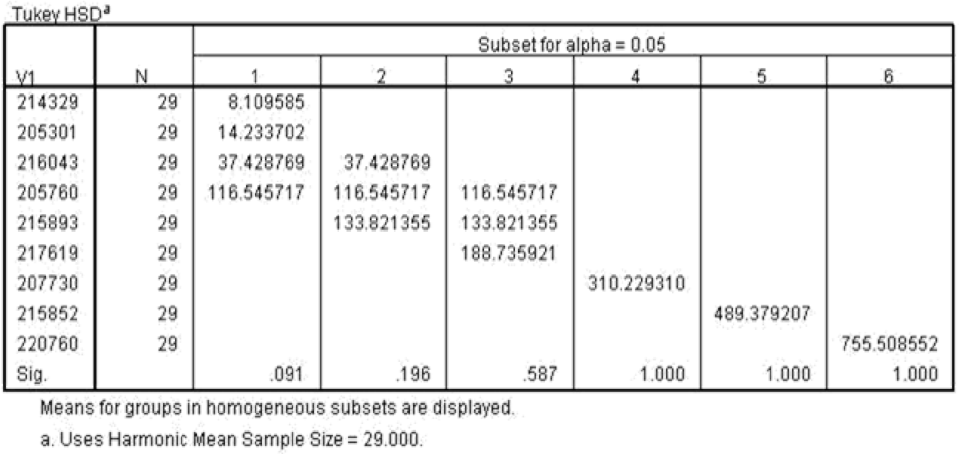


Fig: 10. Cross verification with SPSS package for high expression profile

This bar diagram actually indicates that the mean values of the microarray data corresponding to the probe sets in the x-axis. These are nothing but probes of the gene that the user has inputted. In this case we checked out for OGG1, which is a BER gene. The mean values exactly coincide with the mean values obtained from the output of the high expression profile (see Fig: 9 y axis).



Tukey HSD[a]

V1	N	Subset for alpha = 0.05					
		1	2	3	4	5	6
214329	29	8.109585					
205301	29	14.233702					
216043	29	37.428769	37.428769				
205760	29	116.545717	116.545717	116.545717			
215893	29		133.821355	133.821355			
217619	29			188.735921			
207730	29				310.229310		
215852	29					489.379207	
220760	29						755.508552
Sig.		.091	.196	.587	1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 29.000.

Fig: 11. Cross verification of expression level significance through SPSS package

This was actually verified for the OGG1 gene for the expression profile that was constructed. The graph that was constructed uses Tukey test if in case ANOVA's hypothesis fails. This helps to identify the difference between the probe sets so as the topmost would be the highest expressed probe. The plot is constructed in such a way that the order of expression level is in the descending order. That is the plot is plotted from the highly expressed probe to the lowly expressed probe. Here in SPSS package one could observe that the highly expressed probe is the same in both of the cases. As the significance value 1 is said to be most significant. If it is nearby 1 then it can be considered. But those nearing 0 are lowly expressed. Thus the results obtained were efficient when verified with SPSS.

The Phase III:

This is the most important phase which depends on both the phase I and phase II. In this phase all the possible protein structures are retrieved retrieves all the possible protein structures that are present or that are found till date. All possible information with respect to the structures like active site, metal binding site, metal ions are listed for the possible structures of the protein. This information is very important to the user as they can further proceed with drug interaction studies. The functions carried out by the genes are also listed along with the ligands that are responsible to switch on or off the protein. The similarities for these ligands are listed on the basis of Tanimoto index and similarity at the range greater than or equal to 95%. Since more than one similar ligand shares the same IUPAC name, the ligands are being tabulated bases upon their molecular formula and smile notation. The molecular descriptor with respect to rule of 5 (Lipinski rule) is also indicated.

VERIFICATION:

The interactions of the ligands and the similar ligands with the protein (1YQK) are analysed through docking. The nucleotide sequence of the probe was converted into amino acid sequence and thus obtained 6 different possible frames of sequences. The results for OGG1 gene shows that the 220760_x_at is the probe that got highly expressed and its nucleotide sequence is TTCGAGACCAGGCTGGCCAACAGGG. For this nucleotide sequence, the 6 possible amino acid sequences were FETRLANT, SRPGWPTR, RDQAGQH, PCWPAWSR, RVGQPGLE and VLASLVS. Among them the second frame SRPGWPTR was found to be the site of action in the 1YQK because this frame was found in the protein structure interconnected with each other forming a pocket. The complex structure of 1YQK also had interaction with these residues. Thus this frame was chosen. Its ligand was identified as Glycerol. The docking experiment was conducted between Glycerol and its similar ligands. The docking experiment was done in order to show the interactions of the ligands and the similar ligands with the 1YQK take place with the same nearby residues in their binding pocket. The docking experiment was done with the help of AutoDock4.2 tool.

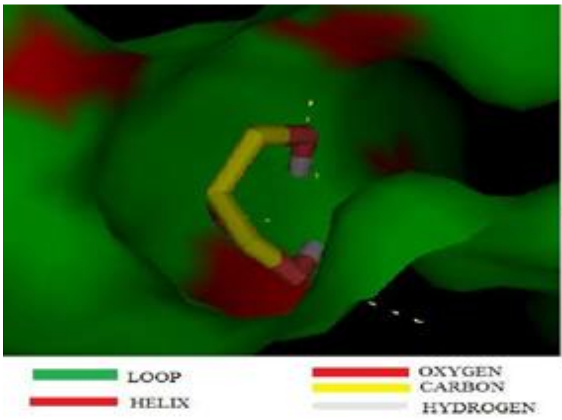


Fig: 12. Docking interaction for original ligand with 1YQK

LOOPS
OXYGEN
CARBON
HYDROGEN

Fig: 13. Docking interaction for similarity ligand 1 with 1YQK

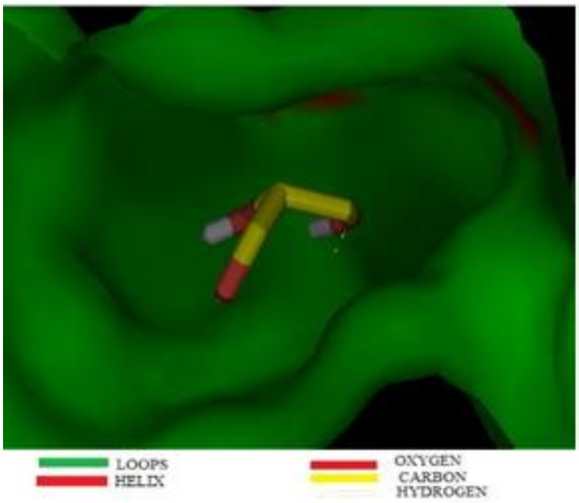


Fig: 14. Docking interaction for similarity ligand 2 with 1YQK

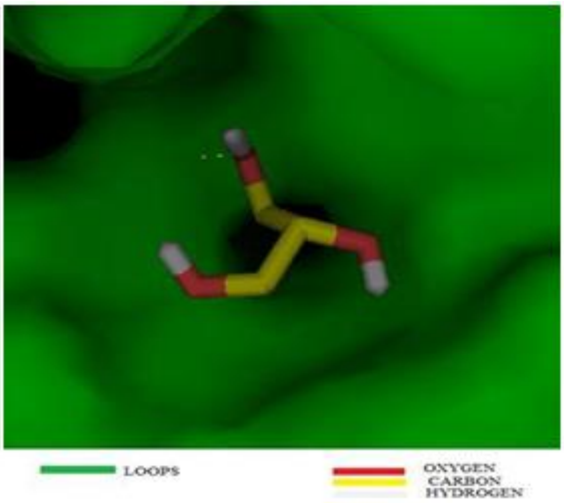
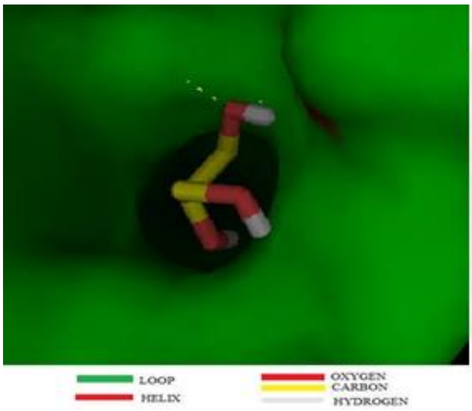


Fig: 15. Docking interaction for similarity ligand 3 with 1YQK



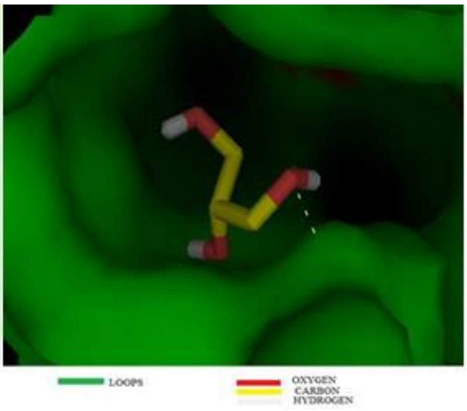
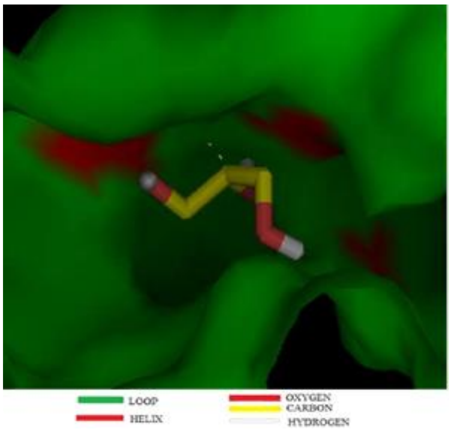


Fig: 16. Docking interaction for similarity L ligand 4 with 1YQK

Fig: 17. Docking interaction for similarity ligand 5 with 1YQK



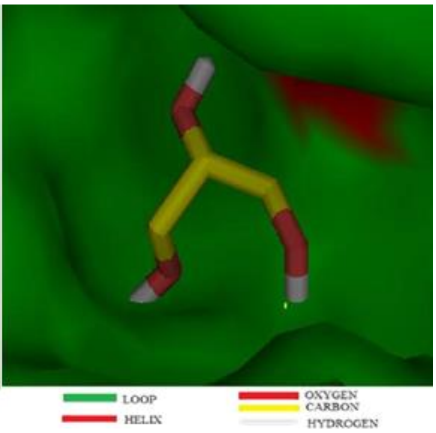


Fig: 18. Docking Interaction for similarity ligand 6 with 1YQK

Fig: 19. Docking interaction for similarity ligand 7 with 1yQK

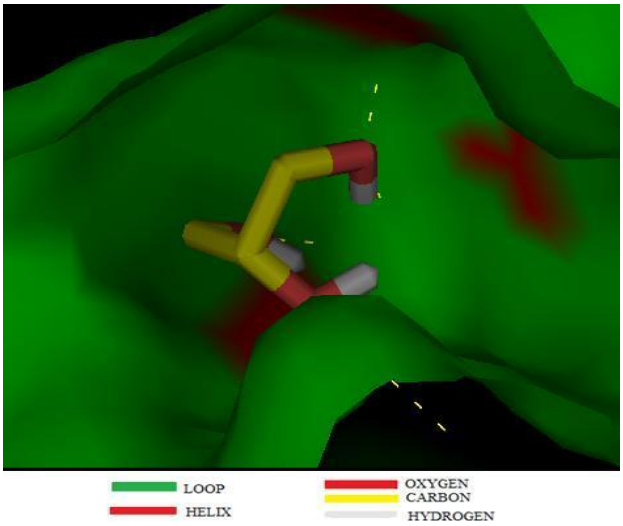


Fig: 20. Docking interaction for similarity ligand 8 with 1YQK

Ligands	SER	ARG	PRO	GLN	TRP	ALA	THR	LEU
Complexed	3.11	-	3.20	-	-	-	-	2.67,1.97,1.95
Similar1	3.20	-	-	3.58,2.94	3.51	-	-	2.91,2.89,2.61
Similar2	-	-	-	2.67	-	-	-	-
Similar3	3.35	-	-	-	-	-	-	-
Similar4	-	3.23	-	-	-	-	2.88	-
Similar5	3.12	-	-	-	-	-	-	-
Similar6	3.34	-	-	-	-	3.44	-	-
Similar7	-	-	-	3.15	-	-	-	-
Similar8	3.26	-	3.41	-	-	-	-	2.75,2.70,2.80

Table. 1: Interaction range between the molecules in the binding pocket

Thus the docking interactions clearly showed that their binding pocket and their nearby residues were the same. There interaction range was almost within 4 Å. Apart from docking interactions there physico-chemical properties were analyzed and were proved to be the same. Their binding residues were checked and found to coincide with the second frame exactly.

CONCLUSION

Microarray data deals with the gene expression. Thus it helps us to identify the portion of the gene that actually expresses well in a particular condition. That portion of the gene could be taken as a domain or the site of action. The methodology helps in identifying this domain by getting the microarray data from the user or from the EBI web services. The documented library is being made use of to retrieve information regarding the proteins and the function with respect to the gene inputted. It uses comparative approach to generate the cluster plot. The images are in the form of JPEG so that the user could download the image. The understanding of the similar functional proteins produced by the genes also helps the user to indirectly understand the concept of alternative splicing mechanism which makes one gene one protein hypothesis to many protein hypothesis. The results help the end user to reduce this time scale from searching out the chemical molecular libraries for similar functional ligands. The use of similarity approaches are done rather than random approach due to its higher accuracy. This information could be widely used in the field of Drug design to identify lead compounds. The percentage criterion set is 95% and the respective tanimoto index has been ranged from 0.95 to 1, which helps in identifying the similar ligands containing the same functional groups as such of the original ligand which activates the proteins. In future we plan in extending this methodology for the whole of class of human genes apart from the BER genes of the DNA repair mechanisms.

ACKNOWLEDGEMENT

We thank SASTRA University for providing us with a wonderful infrastructure to successfully complete our research.

REFERENCES

[1] NP Singh, MT McCoy, RR Tice, EL Schnider (1988). A simple technique for quantitation of low levels of DNA damage in individual cells. Experimental cell research. 175(1): 184 – 191.

[2] T Lindahl, P Karran and RD Wood. (1997). DNA excision repair pathways. Genetics and development. 7:158-169.

[3] Y Liu, R Prasad, WA Beard, PS Kedar, EW Hou, DD Shock, SH Wilson. (2007). Coordination of Steps in Single-nucleotide Base Excision Repair Mediated by Apurinic/Apyrimidinic Endonuclease 1 and DNA Polymerase beta, The Journal of Biological Biochemistry. 282 (18): 13532–13541

[4] I Muegge, YC Martin (1999). A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. J. Med. Chem. 42(5): 791 – 804

[5] JC Fromme, GL Verdine (2004). Base Excision Repair. Advances in Protein Chemistry. 69. 1 – 41.

[6] R Edgar, M Domrachev, AE Lash. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucl.Acids Res. 30 (1):207-210.

[7] J Garbmeier, A Rudolph. (2002). Techniques of Cluster Algorithms in Data Mining. Data mining and Knowledge discovery. 6(4). 303 – 360.

[8] DC Hoaglin, RE Welsh (2012). The Hat Matrix in Regression and ANOVA. The American Statistician. 32(1): 17 – 22

[9] BH Marcus, N Oven. (2006). Motivational Readiness, Self-Efficacy and Decision-Making for Exercise. Journal of Applied Social Psychology. 22(1): 3 – 16

[10] RA Irizzary, BM Bolstad, F Collin, LM Cope, B Hobbs, TP Speed. (2003). Summaries of Affymetrix GeneChip probe level data. Nucl. Acids Res. 31(4): 15.

[11] D Butina. (1999). Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.*, *39* (4), pp 747–750

[12] The PHP: http://www.php.net/ . (Accessed on November 2012)

[13] A Kouranov, L Xie, J de la Cruz, L Chen, J Westbrook, PE Bourne, HM Berman. (2005). The RCSB PDB information portal for structural genomics. 34(1): 302 – 305

[14] CA Lipinski. (2004). Lead- and drug-like compounds: the rule-of-five revolution. Drug Discovery Today: Technologies. 4(1): 337 – 341

[15] NH Nie. SPSS statistical package for the social sciences. Mc-Graw Hill. New York. 1975