Research Journal of Recent Sciences _____ ISSN 2277-2502
Vol. **6(7),** 63-70, July **(2017)** *Res. J. Recent Sci.*

Review Paper

Genome wide screening and analysis of *Homo sapiens* genes and proteins associated with schizophrenia

Sushma Rani Martha[1,2]*, Dibyashree Mallik[1] and Manorama Patri[1]
[1]Department of Life Sciences, Ravenshaw University, Cuttack, India
[2]Department of Bioinformatics, Odisha University of Agriculture and Technology, Bhubaneswar, India
sushma.martha@gmail.com

Available online at: www.isca.in, www.isca.me
Received 1st May 2017, revised 20th June 2017, accepted 1st July 2017

Abstract

Implementing traditional methods of browsing scientific information in literature databases like PUBMED to browse the molecular basis of a complex brain disorder like schizophrenia revealed the involvement of hundreds of gene with the disorder. It urged for the necessity to adopt some specialized experimental design including more relevant and reliable data mining technologies and methodologies to screen out the key player genes and proteins involved with our targeted disorder. After searching for all possible available molecular data for schizophrenia in case of Homo sapiens not less than 400 genes were found to be reported in about 900 different studies through GWAS. Various types of further analysis were then carried out on this gene set to filter the exact genes and proteins involved in the disorder based on their physicochemical properties, chromosomal localization, pathway analysis, involvement in biological processes, cellular localization, drug association studies and disease association studies. After all tedious observations and analysis interestingly it is revealed that the human chromosome No. 22 is highly enriched with schizophrenia associated genes, most of the genes are linked with more than one disorder along with schizophrenia, most of the proteins are membrane proteins and very less proteins are available with drugs approved for the disorder.

Keywords: Schizophrenia, Data mining, Drug Association, Disease Association, Chromosomal Enrichment.

Introduction

Schizophrenia is a neurological disorder fairly known as cancer of the brain, has remained a big mystery at the molecular level even after research of several years by number of researchers. Since the exact genetic cause and method of its eradication have not yet been pinpointed the available treatments can only suppress but cannot cure the disorder.

Genetics, early environment like prenatal stressors, neurobiology, psychological and social processes appear to be important causative factors of the disorder where as some recreational and prescription drugs appear to cause or worse the symptoms. Diagnosis of the disease may be either on symptomatic or on physiological or on genetic basis[1,2].

Huge number of genes and proteins are found to be responsible and related with this disease. So, retrieving relevant publications through the PubMed search engine and creating gene lists is not only time-consuming, but also prone to errors[6].

Therefore our study is targeted to carry out data mining of all possible molecular data on Schizophrenia in case of *Homo sapiens* and then screen out the major key player genes and proteins involved in the disorder by implementing best possible methodology with best possible experimental design.

Methodology

After extensive review of literature we opted six databases which provide annotated molecular level information. These are GWAS, GLAD4U, PharmGKB, DisGeNET, GenAtlas and Uniprot. First two databases provide information from the publications available at PUBMED based on their respective internal cutoff parameters. Next three databases are databases of genomic disorders. And last one is the database of protein sequences of a number of organisms including human.

GWAS: GWAS is a huge repository of all published genome wide association studies. National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI) jointly host the database[3-5]. We collected genes taking P value $\leq 5*10^{-8}$ and schizophrenia as reported trait.

GLAD4U: Gene List Derived for You is a user interface that accepts any valid queries for PubMed, and its output page displays the ranked gene list and information associated with each gene[6]. We extracted only those genes from this database that are having a cutoff score >1.

PharmGKB: PharmGKB is a pharmacogenomics knowledge resource that encompasses clinical information including dosing guidelines and drug labels, potentially clinically actionable

gene-drug associations and genotype-phenotype relationships. PharmGKB collects, curates and disseminates knowledge about the impact of human genetic variation on drug responses[7,8].

DisGeNET: The DisGeNET database integrates human gene-disease associations (GDAs) from various expert curated databases and text-mining derived associations including Mendelian, complex and environmental diseases[9,10]. From this database we took genes of the searches with cutoff score > = 1.

GenAtlas: The GenAtlas database provides information on the structure, expression and function of genes, gene mutations and their consequences on diseases. GenAtlas is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine[11]. We took only those sequences which are having any valid Uniprot Id.

Uniprot: The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). Across the three institutes more than 100 people are involved in the database curation, software development and support[12,13]. We downloaded all sequences in excel and removed redundancies.

DrugBank: The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed chemical, pharmacological and pharmaceutical drug data with comprehensive drug target information. It contains extensive data on the nomenclature, ontology, chemistry, structure, function, action, pharmacology, pharmacokinetics, metabolism and pharmaceutical properties of both small molecule and large molecule drugs. It contains information on approved, investigational as well as withdrawn drugs[14,15]. We found all the drugs available in this database for Schizophrenia which came out to be 80 in number. Then we collected all the target protein for each of the 80 drugs and compared these drug targets with the previously collected lists of proteins.

String: It is a database for prediction of protein-protein interaction within a list of proteins submitted and provides the result in the form of network showing the number of links arising from each protein assigning a weightage to each link. The interactions include direct (physical) and indirect (functional) associations[16]. We uploaded protein lists collected from above mentioned six databases to construct network using STRING. Then we recorded the number of second order protein linked to each and every individual protein and screened out only those proteins that are having greater than the one third of the highest number of links for each protein list.

GeneCards: This is a database which provides information on the number of diseases associated for any queried protein[18,19]. We collected the diseases associated with all the drug target proteins after which we selected only those proteins that were found to be strongly associated with Schizophrenia as well as with at least ten numbers of diseases including Schizophrenia. Then we compared these proteins with the filtered protein lists generated through STRING.

NCBI Gene: NCBI Gene is the genomic database hosted by National Centre for Biotechnology Information. After merging all the gene lists we noted the chromosomal locations of each gene using this database. Then we computed the chromosomal distribution of these genes by comparing all the cytogenetic locations. We also calculated the chromosomal enrichment value for each chromosome.

Panther: PANTHER (protein analysis through evolutionary relationships) Classification System was designed to classify proteins and their genes in order to facilitate high-throughput analysis. It is a software which facilitates classification system for different proteins and their genes1[17,20]. We found and analysed the Molecular functions, Biological processes and Cellular components of all proteins that are reported in any of the six considered databases, that passed through STRINGS screening process, that are having drugs in the DrugBank and that are also strongly associated with at least ten disorders including Schizophrenia.

ProtParam: Expasy Protparam is a tool which allows the computation of various physical and chemical parameters for a given protein. Parameters like Atomic composition, Isoelectric point, Molecular weight, Instability index, Number of positive and negative residues, Half life period and (GRAVY) grand average of hydropathycity value are computed for one protein sequence at a time[21]. We made all these physicochemical characterizations and recorded all parameters in a single table to make comparative analysis.

Lastly we traced back to extract the number of drugs associated, number of diseases associated and the number of databases associated for the screened out proteins and tried for further final round screening and comparative analysis of the proteins.

Results and discussion

Gene/protein list: It is found that 399 genes from GWAS, 417 genes from GLAD4U, 366 genes from PharmGKB, 259 genes from DisGeNET database, 200 genes from GenAtlas and 142 genes from Uniprot database. 80 number of drugs are enlisted for schizophrenia in the DrugBank that are having 233 number target proteins out of which genes for 115 number of proteins are found to be reported by any of the six databases (Figure-1).

Filtered gene/protein list: After analyzing the gene lists of each database using STRING we got 40 genes from GWAS, 55 genes from GLAD4U, 53 genes from PharmGKB, 32 genes from DisGeNET database, 32 genes from GenAtlas and 31

genes from Uniprot database. So, these are the important proteins knocking out which may cause serious alteration in the normal biological processes as they are found to be functionally linked to large number of other proteins. Number of DrugBank proteins which are strongly associated with at least ten disorders along with schizophrenia is found to be 37. Out of these proteins, 26 numbers of proteins are available in the above filtered protein lists of different databases (Figure-2).

Chromosomal Distribution and Enrichment: Gene list of all databases are merged with each other and that of the genes of DrugBank target proteins to form a dataset of 151genes. Comparing chromosomal locations of all these genes led to the observation that human chromosome No. 7 is having highest i.e. 15 number of genetic loci for Schizophrenia genes whereas Chromosome No.8, 21 and X are having least number of genetic Loci i.e. 1. And also Chromosome No. 18 is completely devoid of any genetic locus. And at the other hand when we calculated Chromosomal enrichment by dividing the number of genetic loci with that of chromosomal length we found that Chromosome No. 22 followed by Chromosome No. 19 is having highest whereas Chromosome No. 8 is having lowest enrichment value (Figure-3 and Table-1).

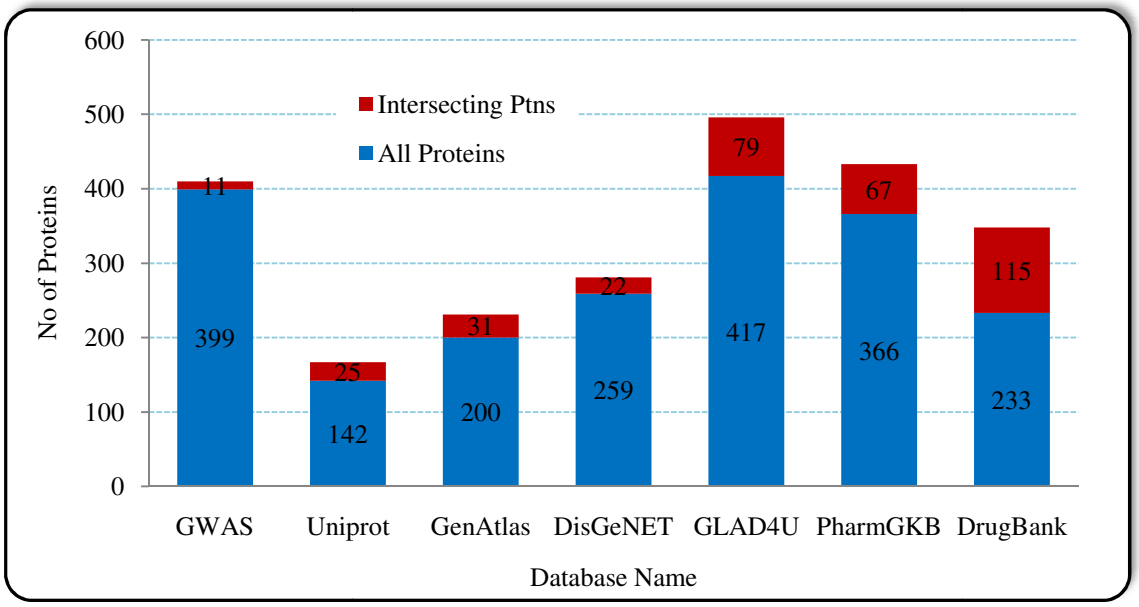

Figure-1: Graph showing proteins collected from different databases.

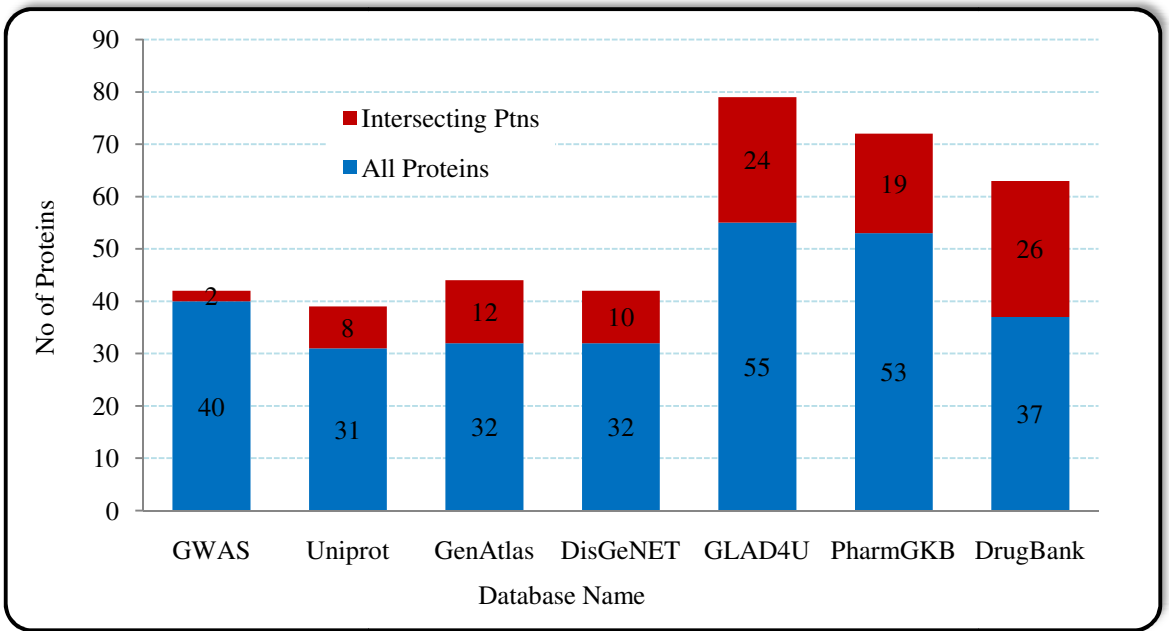

Figure-2: Graph showing proteins filtered by STRING network analysis.

Protein Functional Classification: After carrying out the functional characterization of above found filtered 26 proteins using PANTHER database we found that functionally they are majorly involved in five different activities having either binding activity or catalytic activity or receptor activity or signal transduction activity or transporter activity and they are distributed in the cell by forming part of either cell membrane or organelle membrane or synapse or any cellular macromolecule. We also observed that maximum (35.7%) of all the proteins are having receptor activity and maximum (37.5%) of them are membrane proteins (Figure-4 and Figure-5).

Table-1: Chromosomal distribution and Chromosomal enrichment values.

Chromosome No	Genes/Chr	Chr Length (Mbp)	Enrichment
1	7	248.96	2.811696658
2	10	242.19	4.128989636
3	9	198.3	4.538577912
4	5	190.22	2.62853538
5	12	181.54	6.610113474
6	11	170.81	6.439903987
7	15	159.35	9.413241293
8	1	145.14	0.688989941
9	6	138.4	4.335260116
10	9	133.8	6.726457399
11	10	135.09	7.402472426
12	6	133.28	4.50180072
13	2	114.36	1.748863239
14	2	107.04	1.868460389
15	8	101.99	7.843906265
16	6	90.34	6.641576267
17	8	83.26	9.608455441
18	0	80.37	0
19	7	58.62	11.94131696
20	3	64.44	4.655493482
21	1	46.71	2.140869193
22	9	50.82	17.70956316
X	4	156.04	2.56344527
Y	1	57.23	1.747335314

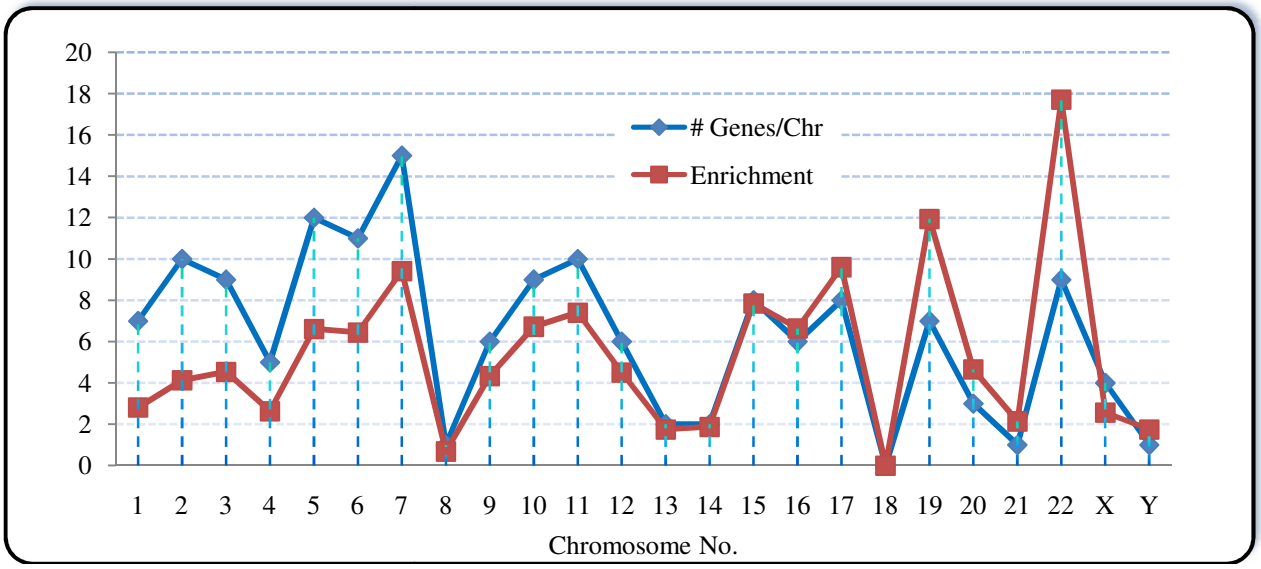

Figure-3: Graph of Chromosomal distribution and enrichment.

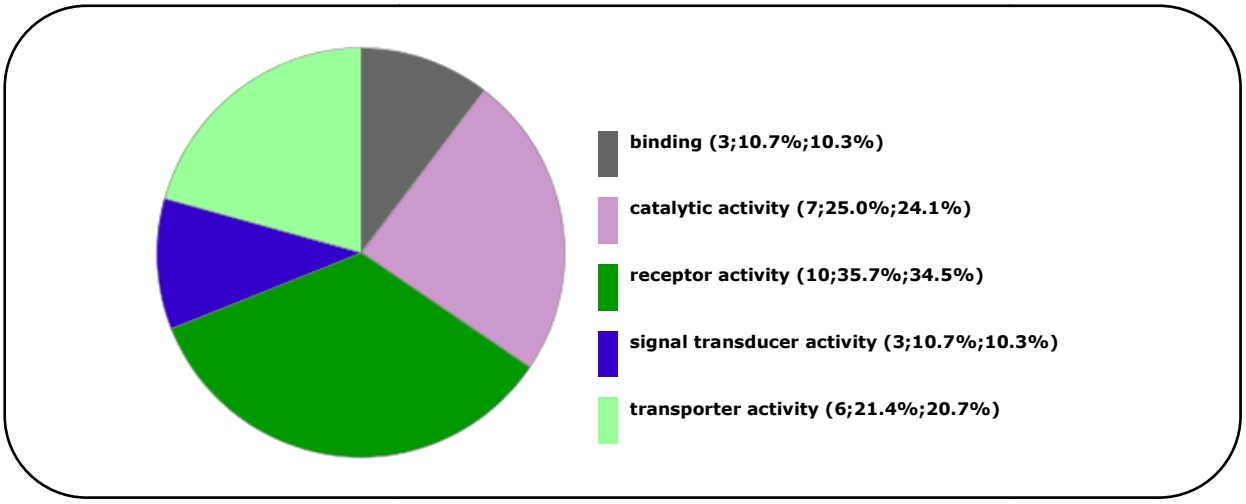

Figure-4: Pie chart showing Molecular functional classification of 26 proteins.

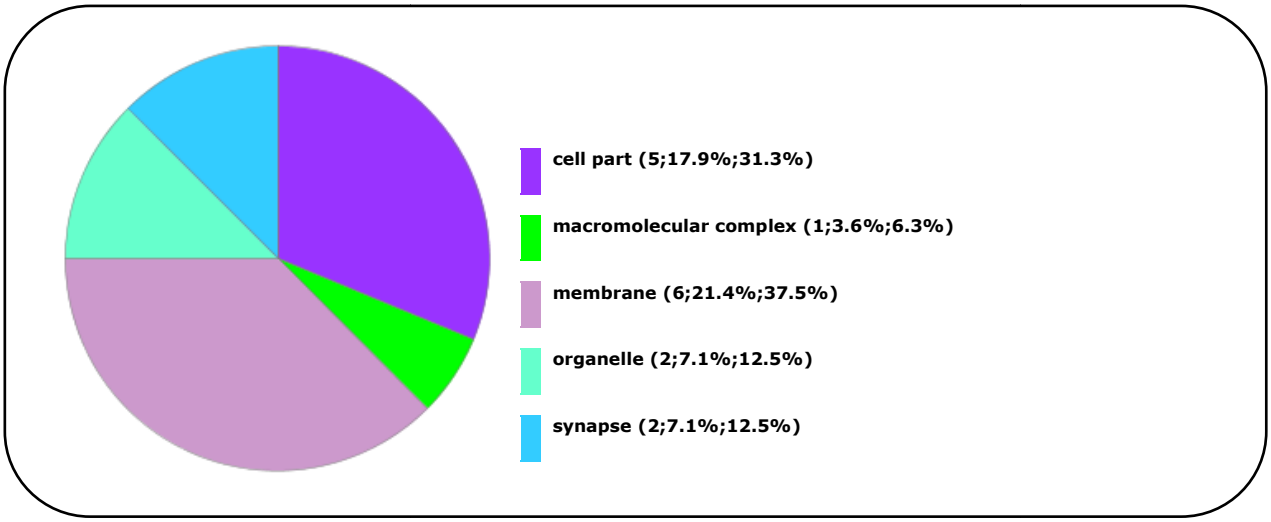

Figure-5: Pie chart showing Cellular localization of 26 proteins.

Primary Sequence Analysis: Comparative analysis of different parameters obtained using ProtParam led to the findings that 14 out of 26 proteins are highly unstable based on Instability Indexes which are above the threshold of 40 and 12 proteins are having Aliphatic Indexes above 100. We also observed that most of the proteins are hydrophobic in nature due to positive GRAVY score. The Theoretical Isoelectric points can also be observed which gives idea about the pH of the proteins (Table-2).

Association Studies: Tracing back the number of drugs, diseases and databases associated with the 26 filtered proteins led to this observation that DRD2 followed by HTR2A followed by HTR1A followed by DRD1 followed by DRD3 followed by DRD4 followed by other proteins can be concluded as best proteins for the disorder on the basis of drug, disease and database associations (Figure-6).

Table-2: Physicochemical properties of filtered proteins.

Intersecting proteins	Atomic Formula	Amino acids	Mol Wt (kDa)	Theoretical pI	Instability Index	Aliphatic Index	GRAVY
CHRNA4	$C_{3170}H_{4960}N_{834}O_{886}S_{32}$	627	69957.23	6.81	54.16	93.89	-0.014
CNR1	$C_{2399}H_{3788}N_{618}O_{665}S_{29}$	472	52857.95	8.47	39.74	105.74	0.329
CYP2D6	$C_{2532}H_{3951}N_{699}O_{692}S_{16}$	497	55769.45	6.77	44.63	95.15	-0.031
DAO	$C_{1787}H_{2743}N_{491}O_{503}S_{10}$	347	39474.02	6.36	31.68	88.21	-0.317
DRD1	$C_{2234}H_{3504}N_{578}O_{627}S_{25}$	446	49293.39	8.64	37	99.93	0.335
DRD2	$C_{2293}H_{3628}N_{620}O_{617}S_{27}$	443	50619.43	9.55	53.08	98.78	0.06
DRD3	$C_{1998}H_{3163}N_{531}O_{552}S_{24}$	400	44224.76	9.2	43.37	102.55	0.319
DRD4	$C_{2150}H_{3424}N_{612}O_{591}S_{33}$	467	48360.56	8.79	50.77	89.94	0.223
DRD5	$C_{2405}H_{3666}N_{618}O_{674}S_{29}$	477	52951.04	5.23	38.38	93.23	0.244
GAD1	$C_{2983}H_{4660}N_{810}O_{877}S_{31}$	594	66896.58	7.54	36.24	79.65	-0.32
GAD2	$C_{2925}H_{4532}N_{780}O_{844}S_{40}$	585	65411.28	6.45	43.11	79.25	-0.201
GRIN1	$C_{4687}H_{7408}N_{1312}O_{1370}S_{41}$	938	105372.81	9.03	40.69	84.52	-0.264
GRIN2A	$C_{7336}H_{11397}N_{2007}O_{2208}S_{70}$	1464	165282.52	6.67	42.26	76.73	-0.393
GRIN2B	$C_{7371}H_{11421}N_{2023}O_{2232}S_{71}$	1484	166367.24	6.47	48.76	74.55	-0.388
GRM1	$C_{5928}H_{9307}N_{1575}O_{1739}S_{59}$	1194	132357.16	6.27	50.43	89.64	-0.1
HTR1A	$C_{2079}H_{3311}N_{559}O_{579}S_{22}$	422	46106.88	9.13	36.52	100.81	0.195
HTR1B	$C_{1998}H_{3139}N_{503}O_{547}S_{19}$	390	43568.08	8.96	43.51	105.77	0.342
HTR2A	$C_{2367}H_{3742}N_{602}O_{688}S_{30}$	471	52603.17	7.83	39.77	101.83	0.221
HTR2C	$C_{2358}H_{3764}N_{622}O_{633}S_{27}$	458	51821.23	9.13	38.42	114.21	0.309
HTR3A	$C_{2543}H_{3956}N_{652}O_{690}S_{18}$	478	55280.42	7.04	48.92	109.44	0.104
HTR7	$C_{2439}H_{3840}N_{640}O_{654}S_{30}$	479	53555.02	9.09	48.61	101.96	0.243
SLC1A2	$C_{2795}H_{4545}N_{723}O_{796}S_{34}$	574	62104.24	6.09	35.37	116.15	0.434
SLC6A2	$C_{3267}H_{4909}N_{783}O_{840}S_{23}$	617	69332.04	7.18	25.14	106.81	0.468
SLC6A3	$C_{3194}H_{4853}N_{775}O_{847}S_{26}$	620	68494.91	6.46	30.85	105.32	0.499
SLC6A4	$C_{3278}H_{4949}N_{781}O_{879}S_{30}$	630	70324.86	5.89	33.13	100.44	0.422
TH	$C_{2619}H_{4070}N_{730}O_{779}S_{11}$	528	58600.21	5.9	49.58	78.81	-0.365

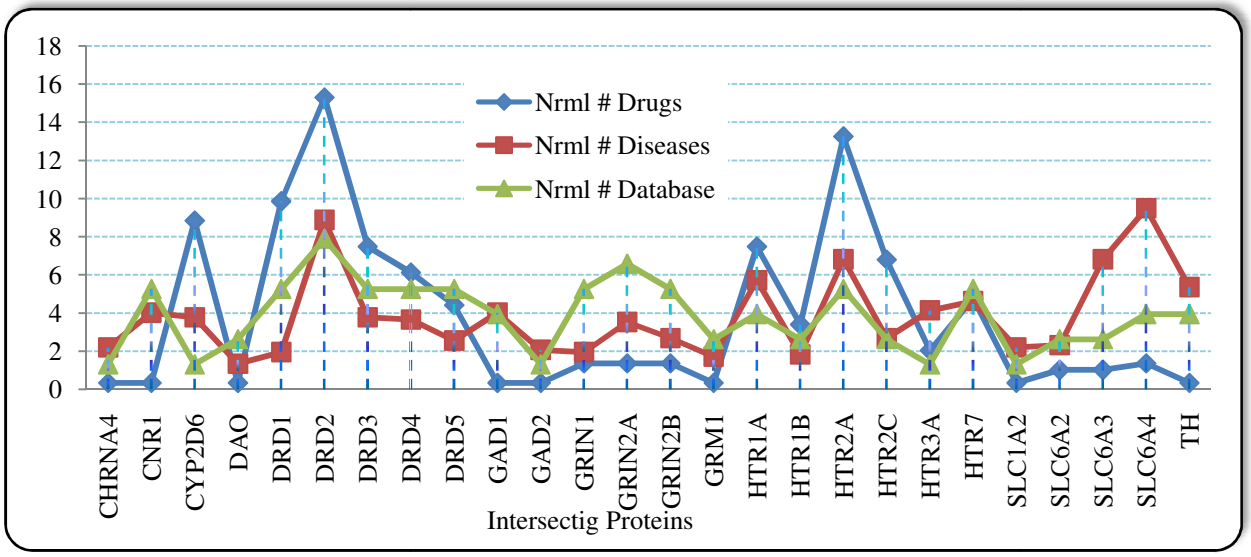

Figure-6: Graph showing Drug, Disease and Database association of filtered proteins.

Conclusion

We can firmly conclude from our observations and findings that DRD2 and HTR2A are the two proteins which are strongly associated with Schizophrenia as we could screen them out from a huge molecular dataset of hundreds of proteins. Our future researches will be more focused on these two Proteins. We may make the secondary and tertiary structural analysis of the proteins. We may thereafter comparatively study the interactions of the proteins with their available drugs, some of which are agonistic some are antagonistic, some are partial agonistic while some others are acting as substrates.

Acknowledgement

Authors are highly acknowledged to the Department of Life Sciences, Ravenshaw University, Cuttack, Odisha, India and Department of Bioinformatics, Orissa University of Agriculture and Technology, Bhubaneswar, Odisha, India for providing Work base to carry out our research.

References

1. Craddock N., O 'Donovan M.C. and Owen M.J. (2005). The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet*., 42(3), 193-204. doi:10.1136/jmg.2005.030718

2. Karam C.S., Ballon J.S., Bivens N.M., Freyberg Z., Girgis R.R., Lizardi-Ortiz J.E., Markx S., Lieberman J. A. and Javitch J.A. (2010). Signaling pathways in schizophrenia: emerging targets and therapeutic strategies. *Trends in Pharmacological Sciences*, 31(8), 381-390. doi: 10.1016/j.tips.2010.05.004.

3. MacArthur J., Bowler E., Cerezo M., Gil L., Hall P., Hastings E., Junkins H., McMahon A., Milano A., Morales J., Pendlington Z., Welter D., Burdett T., Hindorff L., Flicek P., Cunningham F. and Parkinson H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1), D896-D901.

4. Welter D., MacArthur J., Morales J., Burdett T., Hall P., Junkins H., Klemm A., Flicek P., Manolio T., Hindorff L. and Parkinson H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1), D1001-D1006.

5. Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S. and Manolio T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, 106(23), 9362-9367.

6. Jourquin J., Duncan D., Shi Z. and Zhang B. (2012). GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics,* 13(Suppl 8), S20. doi: 10.1186/1471-2164-13-S8-S20

7. McDonagh E.M., Whirl-Carrillo M., Altman R.B. and Klein T.E. (2015). Enabling the Curation of Your Pharmacogenetic Study. *Clinical Pharmacology & Therapeutics,* 97(2), 116-119. doi: 10.1002/cpt.15

8. Kevin H.J., Jeffrey R.B., Katrin S., Daniel J.M., Yuan J., Susan G.L., Steven L.J., Rebecca L.G., Dana L.C., Adrian L.L., Todd C.S., Stuart A.S., Julia C.S., Teri E.K., Kelly E.C. and Andrea G. (2015). Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and CYP2C19 Genotypes and Dosing of Selective Serotonin Reuptake Inhibitors. *Clin Pharmacol Ther.*, 98(2), 127-134. doi: 10.1002/cpt.147.

9. Bauer-Mehren A., Bundschus M., Rautschka M., Mayer M.A., Sanz F. and Furlong L.I. (2011). Gene-Disease

Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases. *PLoS ONE,* 6(6), e20284. https://doi.org/10.1371/journal.pone.0020284

10. Bauer-Mehren A., Bundschus M., Rautschka M., Mayer M.A., Sanz F. and Furlong L. I. (2010). DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics,* 26(22), 2924-2926. doi:10.1093/bioinformatics/btq538

11. Frezal J. (1998). Genatlas database, genes and development defects. *C. R. Acad Sci III,* 321(10), 805-817.

12. Boutet E., Lieberherr D., Tognolli M., Schneider M., Bansal P., Bridge A.J., Poux S., Bougueleret L. and Xenarios I. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol Biol*., 1374, 23-54. doi: 10.1007/978-1-4939-3167-5_2.

13. Alpi E., Griss J., Wilter A., Silva S., Bely B., Antunes R., Zellner H., Daniel Ríos, Claire O'Donovan, Juan Antonio Vizcaíno, and Martin Maria J. (2014). Analysis of the tryptic search space in UniProt databases. *Proteomics*, 15(1), 48-57. doi: 10.1002/pmic.201400227

14. Knox C., Law V., Jewison T., Liu P., Ly S., Frolkis A., Pon A., Banco K., Mak C., Neveu V., Djoumbou Y., Eisner R., Guo A.C. and Wishart D.S. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.,* 39(suppl_1), D1035-D1041. doi: 10.1093/nar/gkq1126.

15. Law V., Knox C., Djoumbou Y., Jewison T., Guo A.C., Liu Y., Maciejewski A., Arndt D., Wilson M., Neveu V., Tang A., Gabriel G., Ly C., Adamjee S., Dame Z.T., Han B., Zhou Y. and Wishart D.S. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*., 42(D1), D1091-D1097. Doi: 10.1093/nar/gkt1068.

16. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K.P., Kuhn M., Bork P., Jensen L.J. and Von Mering C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(D1), D447-D452. doi: 10.1093/nar/gku1003.

17. Mi H., Dong Q., Muruganujan A., Gaudet P., Lewis S. and Thomas P.D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.,* 38(suppl_1), D204-D210. doi: 10.1093/nar/gkp1019.

18. Rebhan M., Chalifa-Caspi V., Prilusky J. and Lancet D. (1997). GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.,* 13(4), 163. doi:10.1016/S0168-9525(97)01103-7

19. Stelzer G., Dalah I., Iny Stein T., Satanower Y., Rosen N., Nativ N., Oz-Levi D., Olender T., Belinky F., Bahir I., Krug H., Perco P., Mayer B., Kolker E., Safran M. and Lancet D. (2011). In-silico Human Genomics with GeneCards. *Human Genomics*, 5(6), 709-717. doi:10.1186/1479-7364-5-6-709.

20. Mi H., Muruganujan A. and Thomas P.D. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*., 41(D1), D377-D386. doi: 10.1093/nar/gks1118.

21. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. and Bairoch A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. *The Proteomics Protocols Handbook, Humana Press,* 571-607.