COMPUTATIONAL ANALYSIS OF GENE PRIORITIZATION APPROACHES FOR SCHIZOPHRENIA

NeemaTufchi[*], Srashti Chaudhary[**], Bhasker Pant[***], Somya Sinha[****],PriyanshiPundir[**], Kumud Pant[*****]. [1]

[*]Department of Obstetrician and Gynaecology, Sir Ganga Ram Hospital, New Delhi-110060, E-mail ID-
[**]Department of Life Sciences, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India ,
[***]Department of Computer Science, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India.
[****]Department of Biotechnology, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India,
[*****]Department of Biotechnology, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India

ABSTRACT
Biological analysis of the genes helps in predicting the chemical process, but handling of data is very difficult because of their huge size, assorted natureand access time overheads. Disorders like schizophrenia can be diagnosed effectively by finding the most relevant genes which are highly associated with the disorder from the number of candidate genes. Traditional method for gene analysis includes single nucleotide polymorphism (SNP) finding, gene mutation analysis and many in-vitro techniques having many limitations like labor intensive, long-lasting, high cost, and inadequate knowledge of genetic materials. To address such issues computational methods are used which have many advantages over traditional method such as cost effective, time saving, pertinent testing as well as validating tools, ample of primary data etc.The current research emphasizes on computational methods for gene prioritization. Gene prioritization was done with the help of text mining approaches for schizophrenia disorder. Fortext mining Pubtator tool was used for gene prioritization. Pubtator prioritized nine genes (DISC1, COMT, NRG1, DTNBP1, DRD3, DAOA, RGS4, GRM3, and PRODH) which are associated with schizophrenia. Further in-vitro studies on these genes may reveal the actual cause behind the disorder.

KEYWORDs: Text mining, schizophrenia, genes, *in-silico*, gene prioritization.

MSC: 62P10

RESUMEN
El análisis biológico de los genes ayuda a predecir el proceso químico, pero el manejo de los datos es muy difícil debido a su gran tamaño, naturaleza variada y sobrecarga de tiempo de acceso. Trastornos como la esquizofrenia se pueden diagnosticar de manera efectiva encontrando los genes más relevantes que están altamente asociados con el trastorno a partir del número de genes candidatos. El método tradicional para el análisis de genes incluye la búsqueda de polimorfismos de nucleótido único (SNP), el análisis de mutaciones de genes y muchas técnicas in vitro que tienen muchas limitaciones, como el conocimiento intensivo, duradero, de alto costo y inadecuado de los materiales genéticos. Para abordar estos problemas, se utilizan métodos computacionales que tienen muchas ventajas sobre los métodos tradicionales, como la rentabilidad, el ahorro de tiempo, las pruebas pertinentes y las herramientas de validación, una gran cantidad de datos primarios, etc. La investigación actual hace hincapié en los métodos computacionales para la priorización de genes. La priorización de genes se realizó con la ayuda de enfoques de minería de texto para el trastorno de esquizofrenia. Para la minería de textos, se utilizó la herramienta Pubtator para la priorización de genes. Pubtator priorizó nueve genes (DISC1, COMT, NRG1, DTNBP1, DRD3, DAOA, RGS4, GRM3 y PRODH) que están asociados con la esquizofrenia. Más estudios in vitro sobre estos genes pueden revelar la causa real detrás del trastorno.

PALABRAS CLAVE: minería de textos, esquizofrenia, genes, in-sillico, priorización de genes

1. INTRODUCTION

Schizophrenia is a severe neurological disorder in which people are not able to distinguish between reality and fiction thus interfering with the person's thinking ability and their decision making power Riba et.al [20]. Schizophrenia affects approximately 24 million people with the prevalence rate of 1 in 300 persons

[1] pant.kumud@gmail.com (corresponding author)

ranging from 20-28 years in male population and 26-32 for female population, in a report by WHO and Picchioni et al. and Castle D et al. (4, 17, and 24), thus causing morbidity and loss to the nation as well as to the affected family. The occurrence rate of the disorder is predicted to be ten times higher with lifetime prevalence between 0.5% to 1% according to American Psychiatric Association (22). The main reason behind the pathophysiology of schizophrenia is still unknown thus imply the heterogeneous nature of the disorder, having multiple aetiological factors contributing to the disorder by Kahn et al. (13). The genetics is the strongest reason in developing schizophrenia other factors include environmental and physiological by Hallmayer (9). The risk of developing schizophrenia in monozygotic twins is 48% and patients having schizophrenic parents are 13% by O'Reilly (16). The risk of developing Schizophrenia in families is shown in figure 1.

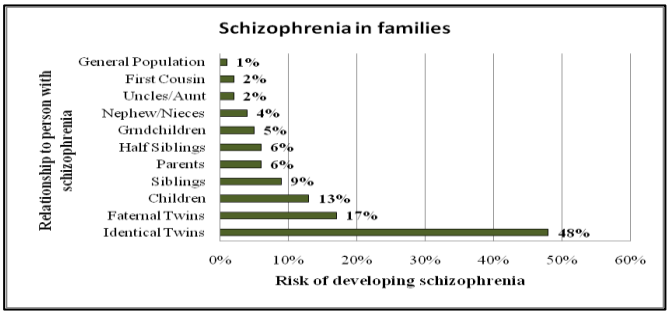

Figure 1: Risk of developing Schizophrenia in families

A lot of research is going on, in predicting the actual reason behind this disorder. There are methods which are generally based on the DNA markers to predict the presence or absence of disease by Henriksen et al. (10). In some studies, DNA variations in or near a candidate gene is considered for the prediction of disease by Salleh (21).

There are many genes which contribute towards schizophrenia and continuous research is going on, but it constantly runs with non-replication of the finding. The reason behind intricacy in gene finding is the heterogeneity, lack of biological marker and complexity of the phenotype by Faraone and Larsson (7). As schizophrenia is a non-mendelian disorder thus it entails the combined effects of several genes, conferring an increase in susceptibility of the disorder by Salleh (21).

Genes have the information regarding protein production for metabolic activities and they are sometimes cited as "the cook book of protein" production. The nucleotide bases present in a gene may get changed due to factors like environmental changes or inheritance by Blazerand Hernandez (2). The mutated nucleotide may either create a faulty protein or no protein. Genetic disorders which rise from inappropriate protein production are categorized into complex disorders or Mendelian disorders by Blazer and Hernandez (2). Mendelian disorder involves the mutation of single gene whereas complex disorder involves the mutation of many genes by Chial (6). Overlapping symptoms, variability in the phenotypic traits and other factors creates many challenges in finding cause behind the complex disorder by Raj and Sreeja (19).

Gene identification is an important step in finding the cause, treatment strategies and prognosis of disorder from the symptoms. Genes need to be prioritized on the basis of their association with the disorder, the process is known as gene prioritization by Bromberg (3). The methods of gene prioritization include wet lab technologies or positional cloning, which are time consuming, expensive and with less precision by Masoudi-Nejad and Meshkin (14). These limitations of gene prioritization methods can be trounced by applying *in-silico* gene prioritization approaches by Chen et al. (5).

For a complex disorder like schizophrenia computational gene prioritization is important in which candidate gene is taken as input and several algorithms are applied to get the output of ranked genes. The top genes in output imply that they are highly responsible for the disorder. Seed genes (small list of genes) can also be uploaded as input, which are associated with the disorder thus seed genes which are similar with the candidate gens are also linked with the disorder by Jia and Zhao (12). Gene prioritization can be done using various methods and tools. A comprehensive literature survey was done which showed that there are several computational methods for gene prioritization methods and data sources used by them are critical.

2. MATERIALS AND METHODS

There are four categories for computational gene prioritization methods which include hybrid strategies, machine learning methods, text mining and network based methods (NBM) by Zolotareva and Kleine (26). Text mining is an approach in which information is retrieved from the published scientific literature to find genes associated with any complex disorder by Zhu et al. (25). NBM uses biological data as a network and applies graph mining techniques to rank different genes Zolotareva and Kleine (26). Machine learning methods are used for studying new algorithms of gene prioritization by Nicholls et al. (15). Hybrid method is an amalgamation of any of these methods. Although these methods are extensively used for gene

prioritization, there is significant increase in the performance by combination of multiple approaches by Chen et al. (5).

Text mining approach

Text mining is a technique to retrieve the data from scientific published data and can be described as the process of exploring huge collections of scientific data to engender new information. The text mining approaches include information retrieval, clustering, document classification, identification of data trends which can be used for ranking of genes by Zhu et al. (25). Keywords used for extracting data from the scientific literature are very crucial for the gene prioritization. There are some metrices which can be used for text mining i.e., Pearson's correlation, cosine similarity, semantic analysis, information content, Jaccard similarity etc. by Aishwarya and Selvi (1). Cosine similarity computes the product of 2 vectors and finds the cosine angle between the two, given by equation 2.1.

$$\cos\theta = \frac{P.Q}{||P||.||Q||} \quad (2.1)$$

where P, Q are two vectors (non-zero).

In text mining, P and Q can be two different sentences, paragraphs or documents. Jaccard similarity is the ratio between union and intersection of two objects by Aishwarya and Selvi (1), shown in equation 2.2.

$$J(P,Q) = \frac{P \cap Q}{P \cup Q} \quad (2.2)$$

Pearson's correlation can be measured as a linear correlation among two variables and can be defined as the ratio between covariance and product of standard deviations by Aishwarya and Selvi (1), equation 2.3.

$$\rho_{p,q=}\frac{cov(p,q)}{\sigma_p \sigma_q} \quad (2.3)$$

Where

cov(p, q)= covariance of the variables p and q

$\sigma_p \sigma_q$= standard deviation between variables p and q.

In semantic similarity calculation, IC (information content) is used to find application of a particular term. "p" denotes the IC of a term, defined as the log ratio of number of objects annotated by all terms (N) to the p annotated objects (np) by Gan et al. (8), shown in equation 2.4.

$$IC(p) = \ln\frac{N}{np} \quad (2.4)$$

Text mining approaches for gene prioritization uses gene ontology, KEGG, MEDLINE and HPRD as a knowledge resource. Gene ontology is a repository with a search engine for fetching the data as it contains key definitions, gene annotations and documentation of terms by Hinderer et al. (11). Text mining can be done on this data to retrieve the genetic disorder, gene and relationship among them. Though text mining approaches can be used in gene prioritization yet has some disadvantages including data redundancy, inaccessible information because of privacy and license issues, accuracy vagueness etc. by Piro and Di Cunto (18).

In the current study text mining approach has been used for the computational prioritization for genes associated with schizophrenia.

The text mining was done using PubTator, a web-based tool for escalating manual literature curation by Wei et al. (23). PubTator is linked with PubMed and comprises of all the published literature with text mining tools applied to articles irrespective of diseases, genes, species, chemicals and mutations. Keywords used to search genes from scientific literature were "schizophrenia genes", "schizophrenia genes SNPs" and "schizophrenia genes mutations SNPs". After short listing with PubTator, the nine genes which are known to cause schizophrenia were retrieved, the detailed information is shown in table 1.

Table 1: Short listed genes through PubTator

S.No.	Genes	Accession Number	Description
1.	*COMT* (Catechol-O-Methyltransferase)	NC_000022.11	Mammalian enzyme known to be involved in metabolic degradation of catecholamines.
2.	*NRG1* (Neuregulin 1)	NC_000008.11	Signalling molecule which has an important role in the organ system growth.
3.	*GRM3* (Glutamate Metabotropic Receptor 3)	NC_000007.14	Neurotransmitter in the mammalian CNS (Central Nervous system), involved in normal brain functions.
4.	*DAOA* (D-amino acid oxidase activator)	NC_000013.11	Functions as an activator of D-amino acid oxidase and are involved in the breakdown of gliotransmitter D-serine.
5.	*PRODH* (Proline dehydrogenase)	NC_000022.11	Enzyme converting proline into D-1-pyrroline-5-carboxylate.

6.	*DTNBP1* (Dystrobrevin Binding Protein 1)	NC_000006.12	Plays pivotal role in regulating the glutamatergic system.
7.	*DISC1* (Disrupted in schizophrenia 1)	NC_000001.11	Gene responsible for mental illness due to its association with dopamine impairments.
8.	*RGS4* (Regulator of G protein signalling 4)	NC_000001.11	Protein which has a role in modulating signalling through G-protein pathways.
9.	*DRD3* (Dopamine receptor D3)	NC_000003.12	D3 receptor is mediated by G proteins which inhibit adenylyl cyclase.

These genes were selected because they have been extensively studied and there is experimental evidence of these genes associated with the pathophysiology of schizophrenia.

3. RESULTS AND DISCUSSION

A total of nine genes were shortlisted because of their role in pathophysiology of schizophrenia also the number of cited literatures of these genes were very high as compared to other genes as shown in figure 2. The number of studies on *COMT* and schizophrenia is shown as 655 (curated on 15-08-2019) on PubTator by using keywords "COMT, Schizophrenia".
There are 510 studies showing association of *NRG1* gene and schizophrenia in PubTator using keywords "NRG1, Schizophrenia". In case of *GRM3* there are 78 published literatures and the keywords used were "GRM3, Schizophrenia". *DAOA* and *PRODH* has 145 and 61 literatures respectively and keywords used were "DAOA, Schizophrenia" and "Proline dehydrogenase gene, schizophrenia" respectively. In case of *DTNBP1*, *DISC1*, *RGS4* and *DRD3* the result come to be of 314, 700, 113, and 310 respectively.

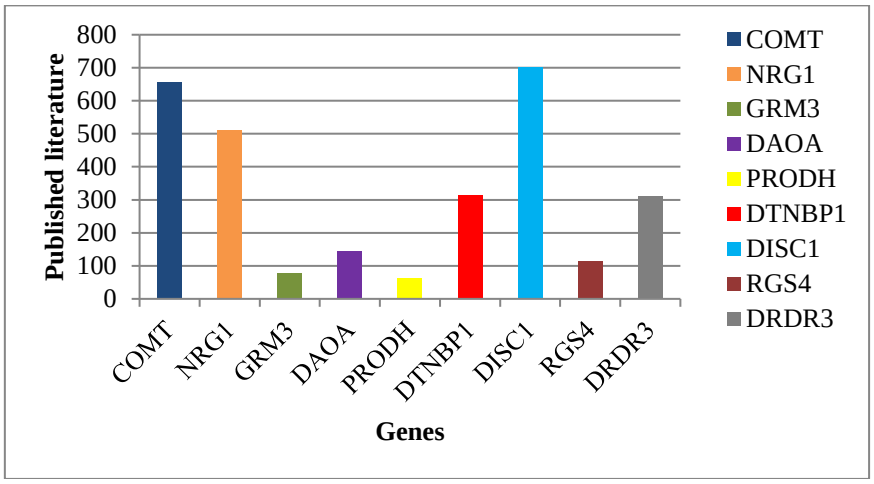

Figure 2: All nine genes with their published literature

4. CONCLUSION

Gene identification is an important for predicting the cause behind any genetic disorder and has been a hot topic due to its broad application in disease prediction. *In-silico* gene prioritization has many advantages over traditional gene prioritization like cost, time and accuracy with the use of many concepts and algorithms to prioritize candidate genes. Also, knowledge-based resources can be used for the prediction and sorting of genes. The databases like NCBI, OMIM and a pathway database KEGG are generally used for the gene prioritization purpose. There are many methods for gene prioritization but in the present study text mining methods have been used against schizophrenia. As schizophrenia is a complex disorder with no known cause, it is essential to prioritize the genes associated with the disorder.According to genome wide association studies (GWAS) there are many genes which have a role in pathophysiology of schizophrenia and to shortlist the gene text mining approach can be beneficial. Text mining approach prioritizes nine genes which may have role in pathophysiology of schizophrenia. Further *in vitro* analysis on these genes can help researchers to better understand the role or functions of these genes in the disorder.

ACKNOWLEDGEMENT: The authors would like to thank department of Biotechnology at Graphic Era deemed to be University for providing the support and computational facility to carry out this work. The authors express the deep sense of gratitude to the Graphic Era University for all the support, assistance, and constant encouragements to carry out this work.

RECEIVED: DECEMBER, 2022.

REVISED: MAY, 2023.

REFERENCES

[1]. AISHWARYA, M. L. and SELVI, K. (2016): An intelligent similarity measure for effective text document clustering. 2016 **International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16),** 1-5.

[2]. BLAZER, D. G. and HERNANDEZ, L. M. (Eds.). (2006): Genes, behaviour, and the social environment: Moving beyond the nature/nurture debate. **National Academies Press**. Available from: https://www.ncbi.nlm.nih.gov/books/NBK19929/ doi: 10.17226/11693.

[3]. BROMBERG, Y. (2013): Disease gene prioritization. **PLoS Comput Biol**, 9(4), e1002902.

[4]. CASTLE, D., WESSELY, S., DER, G. and MURRAY, RM. (1991): The incidence of operationally defined schizophrenia in Camberwell, 1965-84.**The British Journal of Psychiatry,** 159(6):790–794. doi:10.1192/bjp.159.6.790. PMID 1790446. S2CID 41661565

[5]. CHEN, Y., WANG, W., ZHOU, Y., SHIELDS, R., CHANDA, S. K., ELSTON, R. C. and LI, J. (2011): In silico gene prioritization by integrating multiple data sources. **PloS one**, 6(6), e21137.

[6]. CHIAL, H. (2008): Rare genetic disorders: learning about genetic disease through gene mapping, SNPs, and microarray data. **Nature education**, 1, 192.

[7]. FARAONE, S. V. and LARSSON, H. (2019): Genetics of attention deficit hyperactivity disorder. **Molecular psychiatry**, 24, 562-575.

[8]. GAN, M., DOU, X. and JIANG, R. (2013): From ontology to semantic similarity: calculation of ontology-based semantic similarity. **The Scientific World Journal**, Article ID 793091.https://doi.org/10.1155/2013/793091.

[9]. HALLMAYER, J. (2000): The epidemiology of the genetic liability for schizophrenia. **Australian and New Zealand journal of psychiatry**, 34, S47-S55.

[10]. HENRIKSEN, M. G., NORDGAARD J. and JANSSON L. B. (2017): Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. **Front. Hum. Neurosci**, 11:322. doi: 10.3389/fnhum.2017.00322.

[11]. HINDERER III, E. W., FLIGHT, R. M., DUBEY, R., MACLEOD, J. N. and MOSELEY, H. N. (2019): Advances in gene ontology utilization improve statistical power of annotation enrichment. **PloS one**, 14, e0220728.

[12]. JIA, P. and ZHAO, Z. (2014): Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. **Human genetics**, 133, 125-138.

[13]. KAHN, R. S., SOMMER, I. E., MURRAY, R. M., MEYER-LINDENBERG, A., WEINBERGER, D. R., CANNON, T. D., O'DONOVAN, M., CORRELL, C. U., KANE, J. M., VAN OS, J. and INSEL, T. R. (2015): Schizophrenia. **Nature Reviews Disease Primers**, 1(15067). https://doi.org/10.1038/nrdp.2015.67.

[14]. MASOUDI-NEJAD, A. and MESHKIN, A. (2014): Retracted Chapter 2 Gene Prioritization Resources and the Evaluation Method. In Gene Prioritization. **Springer Briefs in Systems Biology. Springer**, **Cham**, pages 9-23.

[15]. NICHOLLS, H. L., JOHN, C. R., WATSON, D. S., MUNROE, P. B., BARNES, M. R. and CABRERA, C. P. (2020): Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. **Frontiers in Genetics,** 11:350. doi: 10.3389/fgene.2020.00350

[16]. O'REILLY, R., TORREY, E. F., RAO, J. and SINGH, S. (2013): Monozygotic twins with early-onset schizophrenia and late-onset bipolar disorder: a case report. **Journal of Medical Case Reports**, *7*, 1-4.

[17]. PICCHIONI, M. M. and MURRAY, R. M. (2007): Schizophrenia.**BMJ,** 335(7610):91–95. doi:10.1136/bmj.39227.616447.BE. PMC 1914490. PMID 17626963.

[18]. PIRO, R. M. and DI CUNTO, F. (2012): Computational approaches to diseasegene prediction: rationale, classification and successes. **The FEBS Journal**, 279(5):678-696.

[19]. RAJ, M. R. and SREEJA, A. (2018): Analysis of computational gene prioritization approaches. **Procedia Computer Science**, 143:395-410.

[20]. RIBA, M., SHARFSTEIN, S. and TASMAN, A. (2005): The American Psychiatric Association. **International Psychiatry**, 2(9):18-20. doi:10.1192/S1749367600007360.
[21]. SALLEH, M. R. (2004): The genetics of schizophrenia. **The Malaysian Journal of Medical Sciences MJMS,** 11(2):3.
[22]. SAMUEL B. GUZE (1994): Diagnostic and statistical manual of mental disorders. 4th ed. (DSM-IV) **American Psychiatric Association, Washington D.C., APA,**873.
[23]. WEI, CH, KAO and HY, LU, Z. (2013): PubTator: a web-based text mining tool for assisting biocuration. **Nucleic Acids Res**., 41(Web Server issue):W518-22. doi: 10.1093/nar/gkt441. Epub 2013 May 22. PMID: 23703206; PMCID: PMC3692066.
[24]. WORLD HEALTH ORGANIZATION (2022): Schizophrenia. retrieved from https://www.who.int/news-room/fact-sheets/detail/schizophrenia.
[25]. ZHU, F., PATUMCHAROENPOL, P., ZHANG, C., YANG, Y., CHAN, J., MEECHAI, A. and SHEN, B. (2013): Biomedical text mining and its applications in cancer research. **Journal of Biomedical Informatics**, 46, 200-211.
[26]. ZOLOTAREVA, O. and KLEINE, M. (2019): A Survey of Gene Prioritization Tools for Mendelian and Complex Human Diseases. **J Integr Bioinform.,** 9, 20180069. doi: 10.1515/jib-2018-0069. PMID: 31494632; PMCID: PMC7074139