International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Precision Mining of Gene-Disease Associations via Frequent Itemset Analysis and Bioinformatics Integration

K.Mary Sudha Rani
Research Scholar,CSE dept., JNTUH Hyderabad,
Assistant Professor,AIML Dept.,
Chaitanya Bharathi of Technology,
Hyderabad.
kmarysudha_cseaiml@cbit.ac.in

Dr.V.Kamakshi Prasad
Professor,CSE dept.,
JNTUH Hyderabad
kamakshiprasad@jntuh.ac.in

Abstract— Biomedical text mining involves the extraction of relevant information from biomedical datasets. It plays a crucial role in genetic research, especially in the development of new drugs where understanding the relationships between genes and diseases is vital. This study introduces a method for generating sets of candidate genes associated with diseases, employing frequent itemset mining for analysis. Genes are ranked based on parameters such as maximum frequent itemset size and gene symbol frequency. This approach aims for precision and efficiency compared to traditional laboratory-based methods, providing highly accurate associations and uncovering novel relationships. Unlike time-consuming laboratory methods, our proposed approach leverages data from the NCBI (National Centre for Biotechnology Information database)via Entrez and utilizes bioinformatics tools like blast for indirect gene associations. Genes exhibiting single nucleotide polymorphisms are identified as indirect genes. The outcomes of this research are anticipated to contribute significantly to biomedical research by offering precise and valuable associations, thereby advancing our understanding of gene-disease relationships..

Keywords- Biomedical Text Mining, Disease-Gene Associations, Frequent Itemset Mining, Indirect Gene Associations, direct Gene Associations.

I. INTRODUCTION

In the realm of modern medicine, the convergence of molecular insights and clinical practice has ushered in a new era where deciphering the intricate relationships between diseases and genes has taken center stage. This burgeoning understanding presents a transformative opportunity, offering a pathway to rectify the fundamental genetic anomalies underlying various diseases[10,11]. By elucidating the gene association's inherent to specific ailments; we navigate closer to correcting these genetic aberrations, a pivotal step in alleviating and potentially eradicating the burden of numerous ailments that afflict humanity.

The identification and comprehension of gene-disease associations stand as a cornerstone in advancing precision medicine[12]. A fundamental aspect in this pursuit is the comprehensive cataloging of associated genes for each disease, meticulously ranked according to multifaceted parameters. This meticulous ranking serves as the bedrock for manufacturing tailored therapeutics and enables a more accurate prognostication of disease trajectories[9]. Our research endeavors aim to fill this critical gap by unraveling the intricate web of relationships between genes and diseases, fostering a roadmap toward more effective treatments and predictive healthcare strategies[8,7].

At the core of our investigative methodology lies the application of frequent itemset mining, principally utilizing the renowned Apriori algorithm. Frequent pattern mining, a technique instrumental in identifying recurring patterns within datasets, particularly frequent itemsets, assumes a pivotal role in our pursuit. Analogous to market basket analysis, this approach endeavors to unearth associations or patterns among genes linked to specific diseases. It assigns crucial metrics such as support and confidence to these associations, mirroring the essence of cross-marketing strategies and offering profound insights into customer behavior in retail settings.

The Apriori Algorithm stands as a cornerstone in our analytical framework, representing an influential tool in mining

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

frequent item sets[13] and formulating Boolean association rules. Employing a methodical "bottom-up" approach, this algorithm systematically extends frequent subsets, incrementally incorporating individual items to uncover latent associations between genes and diseases.

In this paper, we delve into the profound implications of understanding gene-disease associations, delineating the significance of our methodology in elucidating these intricate relationships. Our research endeavors are poised to uncover nuanced insights into disease etiology, thereby fostering a more profound understanding of pathophysiological mechanisms and offering unprecedented opportunities for therapeutic innovation.

Paper objective : Our project's primary goal is to identify every gene linked to a disease and rank them according to a number of criteria that will help with the development of medications and precise forecasting. It is crucial to understand the relationship between genes and disease.

II. RELATED WORK

The methodologies presented by Jae-Yoon Jung et al. [6] and Sune Pletscher-Frankild et al. [1] rely on co-occurrence. However, these approaches exclusively consider abstracts of articles, limiting their ability to extract associations solely present in the main text of the articles. In contrast, Sreekala S et al.'s [3] paper introduces the Hidden Markov Model for identification. This model is coupled with a rule-based Named Entity Recognition approach to identify gene symbols using full-text articles from PubMed, proving more efficient in discovering associations mentioned exclusively in the main text of the literature.

Wu et al. [5] introduced a system for extracting disease-gene associations from biomedical abstracts. They employed a dictionary-based tagger for human genes and diseases, implementing a scoring scheme that considered co-occurrences within and between sentences. This approach successfully extracted a significant portion of manually curated associations with a low false positive rate (0.16%). Additionally, to complement text mining, they developed the DISEASES resource. This resource integrates text mining outcomes with manually curated disease-gene associations, cancer mutation data, and genome-wide association studies. The DISEASES platform, accessible through a web interface, provides text-mining software and associations for download.

DisGeNET, developed by Pi˜nero et al. [2], offers a comprehensive platform focused on understanding the genetic basis of human diseases. With over 380,000 associations between 16,000+ genes and 13,000 diseases, DisGeNET integrates curated databases and text-mined data, covering both Mendelian and complex diseases, including information from animal disease models. Featuring a scoring system based on evidence, DisGeNET provides accessibility through a web interface, a Cytoscape plugin, and a Semantic Web resource. It offers user-friendly data exploration, navigation, and analysis via Cytoscape, facilitating investigations into molecular mechanisms underlying genetic diseases.

Hou et al. [4] proposed two methods to guide gene-disease associations, leveraging proximity relationships between genes and diseases and employing Gene Ontology (GO) term similarity. Their experiments demonstrated that utilizing GO terms outperformed word proximity for associations. This study emphasizes the effectiveness of GO terms as a valuable feature for determining gene-disease associations.

III. MATERIAL AND METHOD

A. Dataset

The National Centre for Biotechnology Information (NCBI) of the United States National Library of Medicine (NLM) created the Entrez database, which is an all-inclusive and integrated platform.. It brings together various databases including PubMed, GenBank, and several other biological databases encompassing genes, proteins, genomes, pathways, and scientific literature. This unified system provides researchers with a singular entry point to access, retrieve, and analyze diverse biological information. Serving as a user-friendly interface, Entrez is instrumental in accessing a broad spectrum of data, including genes, proteins, nucleotide sequences, molecular structures, and biomedical literature. Researchers utilize this database system for tasks like data mining, allowing them to acquire and analyze pertinent information essential for studies in genetics, genomics, molecular biology, and biomedicine.

In this method, the dataset utilized comprises full-text articles centered on genetics sourced from PubMed Central (PMC). To obtain relevant information, the PMC query is tailored to extract articles related to genetics that contain disease names or Medical Subject Headings (MeSH) terms associated with specific diseases within their titles. The chosen diseases for this experiment include Autism Spectrum Disorder, Prostate Cancer, Alzheimer's disease, Bipolar Disorder, and Breast Cancer. This meticulous selection aims to gather specific articles that encompass genetics in relation to these specified diseases for subsequent analysis and research purposes.

B. Tools used

The utilization of specific tools and technologies in computational biology and bioinformatics has significantly enhanced research capabilities. Following tools are used.

1) Anaconda: Anaconda, a comprehensive distribution, simplifies package management and environment configuration. It alleviates dependency conflicts commonly encountered with the pip package manager. Conda ensures compatibility among packages by analyzing the environment before installation, addressing the challenges of managing dependencies in data science projects. It allows the installation of packages from various repositories and aids in creating custom packages using the `conda build` command.

2) Jupyter Notebook: Jupyter Notebook provides an interactive web-based environment for creating documents with code, text, mathematical expressions, plots, and media outputs. The notebook's versatility allows conversion to multiple output formats like HTML, LaTeX, or PDF. Supporting various programming languages through kernels, it fosters collaboration and sharing of computational analyses. JupyterLab, the advanced interface, integrates various tools for a more flexible user experience.

3) NLTK (Natural Language Toolkit): NLTK, is a collection of Python libraries that facilitates a number of tasks related to natural language processing, including parsing, tokenization, tagging, and semantic reasoning. It serves as an educational tool for understanding language processing

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

concepts and aids in building research systems. NLTK's wide adoption in universities and research institutions underlines its significance in teaching and prototyping NLP models.

4) Biopython: Biopython offers a plethora of tools and modules for computational biology and bioinformatics. It assists in sequence representation, files format handling, online database access, and extend functionalities to sequence alignment, population genetics, phylogenetic, and machine learning. This open-source project minimizes code duplication in the domain.

5) FASTA: FASTA, a sequence searching tool, employs local sequence alignment for identifying similarities in nucleotide or amino acid sequences against databases. Its heuristic approach efficiently searches sequences while accounting for word hits and performs optimized searches using a Smith-Waterman algorithm. It's widely used for inferring functional and evolutionary relationships.

6) BLAST: BLAST (Basic Local Alignment Search Tool) revolutionized sequence searching with its heuristic algorithm, significantly faster than traditional alignment methods. Though not guaranteeing optimal alignments, its speed and comparative sensitivity make it essential for quickly identifying sequence similarities. BLAST is pivotal in various bioinformatics research, enabling scientists to explore genetic relationships, protein structures, and more.

These tools and technologies collectively empower researchers in computational biology and bioinformatics, facilitating diverse analyses and discoveries in biological sciences. Their versatility, speed, and ease of use contribute significantly to advancing biological research.

C. *Methodology*

After collecting the articles for each disease, they are converted from PDF files to text files to facilitate text mining. The process involves frequent itemset mining to identify associations between various genes and the respective diseases. Fig. 1. Shows block diagram for analysis of disease associated genes. The steps involved in this method are:

1) Gene Symbol Extraction: This step involves extracting gene symbols or identifiers from the gathered articles. Gene symbols are specific abbreviations or labels assigned to genes, enabling their identification and representation. Techniques such as natural language processing (NLP) or regular expressions may be used to recognize and extract these symbols from the textual content of the articles.

2) Candidate Gene Sets: After extracting gene symbols, candidate gene sets are formed based on these symbols. These sets consist of genes that have been identified and extracted from the articles related to the diseases under consideration. These sets serve as a preliminary pool of genes associated with the diseases, providing the basis for further analysis.

3) Frequent Itemset Mining: Frequent itemset mining involves using algorithms like Apriori or FP-growth to discover patterns or associations among items in a dataset. In this context, the gene symbols extracted earlier form the dataset. The goal here is to identify sets of genes that frequently co-occur or are associated within the dataset of articles[14]. This helps identify patterns of genes that tend to appear together in the context of particular diseases.

4) Frequent Gene Sets: From the results of frequent itemset mining, sets of genes that frequently co-occur or are associated with the diseases are determined. These sets, referred to as frequent gene sets, consist of groups of genes that exhibit strong associations or correlations with the diseases based on their co-occurrence patterns identified in the articles.

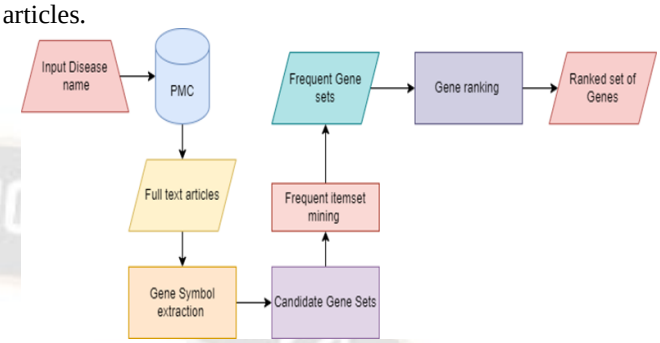

Figure 1. Block Diagram for Analysis of Disease Associated Genes

5) Ranked Set of Genes: Finally, the identified genes are ranked based on specific parameters or criteria. These parameters could include the frequency of occurrence of a gene within the frequent gene sets, the support or confidence level of its association with the diseases, or other relevant factors. Ranking the genes helps prioritize or understand their strengths of association with the diseases, aiding in the selection of potential targets for further research or drug development. Each step contributes to the process of identifying and analyzing gene-disease associations, ultimately providing valuable insights into potential relationships between genes and specific diseases.

D. *Design steps*

As shown in Fig. 2

- The first important step in the design process is data collection from various sources like pubmed, ncbi, genome home reference.
- Next step we have to mine articles collected from the first step.
- Extract gene symbols from the text articles.
- Remove unnecessary gene symbols by using natural language processing tools like nltk.
- Construct a dataset of gene symbols and the article number.
- Apply apriori algorithm to find frequent item sets.
- Rank the given set of genes based on support count and confidence.
- For every gene symbol obtained from the above step, find the gene sequence by querying the entrez database using bio python tools.
- For the gene sequence obtained in the above step use blast tools to find all indirect associations of given gene sequence.
- The result obtained from blast is in the form of XML so we have parsed it to get required data.
- Parse the XML using NCBIXML tool and get all the indirectly associated genes

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

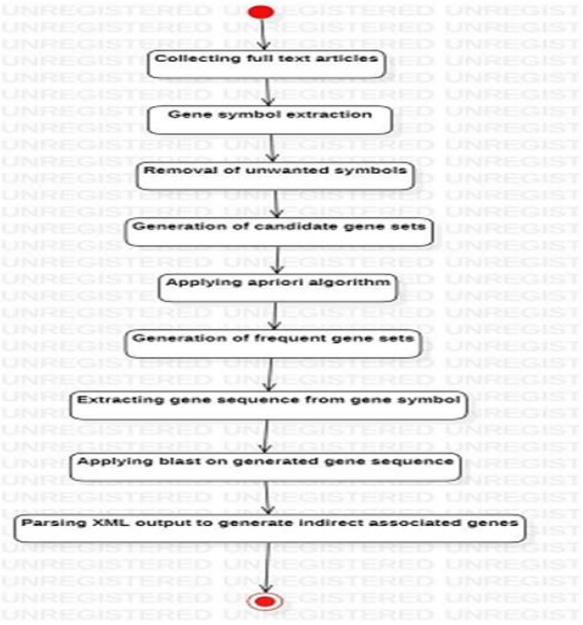

Figure 2. Activity Diagram of analysis of disease associated genes

Algorithm 1 : Algorithm for gene disease association.

Input: Vector C = *<a1, a2, ..an>* of full text PMC articles, d is name of disease
Output: Ranked list of *genes tt = <g1, g2, ..gk> where Score(gi) > Score(gi+1)*

1: *p* = partitions' number.
2: Search the corpus C and extract *D = <a1, a2, ..ax>* where each *ai* is related to the disease d.
3: **for** each article *ai* in - **do**
4: Preprocess *ai* and extract the text that appears between the conclusion and the abstract (ai)
5: **for** each word sequence *wi = <w1, w2, ..wp>* in *ai* **do**
6**:** **if** *wi* corresponds to the gene symbol RE **then**
7: Append *wi* to the candidate gene symbols list *Wi.*
8: **end if**
9: **end for**
10: Remove non-gene-related terms from Wi (dictionary words, disease name abbreviations).
11: **end for**
12: transaction database *B = <ai, Wi>* where candidate gene symbols Wi are items and articles *ai* in *D* are transactions.
13: *L = Partition(B, p) where L = <L1, L2, ..Ly>* is the frequent gene sets .
14: **for** each Li in L **do**
15*:* *tt = tt + Li* where *Li = <g1, g2, ..gr>* and tt (frequently occurring set of gene symbols.)
16: **end for**
17: **for** each *gi* in *tt* **do**
18: Calculate *Score(gi)* .
19: **end for**
20: Sort *tt = <g1, g2, ..gk>* such that *Score(gi) > Score(gi+1).*

IV. RESULTS AND DISCUSSION

The Fig. 3 shows the dataset used to find disease gene associations which consists of record number i.e. article number and set of gene symbols obtained from every article for one particular disease, similar kind of datasets are constructed for every disease. This data set is then used to find disease gene associations.

	record no	gene symbols
0	R1	[ABCA7, AD-, NGS, SORL1, TREM2]
1	R2	[APOE, BELNEU, FTD, R47H, TREM2, VIB]
2	R3	[CIBERER, DAT, IIS-FJD, JAD-170590, SORL1]
3	R4	[TREM2]
4	R5	[PMC5010724, R136Q, R47C, R47H, S31F, SKAT, TR...
5	R6	[ABCA7, AIM, APOE, BDR, CD33, CLU, CONCLUSIONS...
6	R7	[CEA, CHU, CNR-MAJ, CNRS, EOAD, IRIB, SORL1, U...
7	R8	[APP, BACE1, EST, PMC6900319, PSEN1, PSEN2]
8	R9	[ABCA7, APOE, CEA, CHRU, CHU, CNR, CNR-MAJ, CN...
9	R10	[APP, PMC2131721, SNP, SORL1, USA]
10	R11	[AC-MAF, APOE, APOE-, APP, CADD, MAF, PMC55671...
11	R12	[CDE, CIBERNED, CIMA, EOAD, EOD, IDIBAPS, IIB,...
12	R13	[H157Y, R47H, TREM2, USA]
13	R14	[GAB2, PICALM, SNP, SORL1]
14	R15	[A673T, ABCA7, APOE, APP, CNR-MAJ, MAF, PSEN1,...
15	R16	[APOE4, CSF, SNP19, SNP21, SNP21G-, SNP23, SNP...
16	R17	[CTL, GSE63060, GSE63061, MCI]
17	R18	[APP]
18	R19	[ABSTRACT, APP, APP717, CJD, FAD, GSS, OS-2, O...
19	R20	[APP]
20	R21	[APP]
21	R22	[APP, D21S210]
22	R23	[E-4]
23	R24	[AP1, AP2, APP, H2B, SP1, TFIID]

Figure 3. Dataset Collection

We have performed analysis on five different diseases and found direct and indirect associations for each of them. The result of directly associations is the collection of gene symbols. The result of indirect associations is the gene sequences.

A. Direct Associations

Direct associations are the frequent gene sets that are obtained by applying apriori algorithm on mined medical abstracts corresponding to every disease.

Disease name:ALZAMIR DISORDER
Associated set of genes: ['PSEN2', 'PSEN1', 'TREM2', 'SORL1', 'ABCA7']
Disease name: BIPOLAR
Associated set of genes:['HTT', 'HTTLPR']
Disease name: BREAST CANCER
Associated set of genes:['BRCA', 'BRCA1', 'BRCA2']

B. Indirect Associations

Indirect associations are derived for each gene symbol obtained from direct associations. The gene symbol is converted into gene sequence and the blast is applied on the gene sequence to get all indirectly associated genes the output shown here is for single gene SQRL1. Fig. 4 Shows sample of Indirectly associated gene sequences.

Input gene:SQRL1

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Gene sequence extracted:

AGCTACGTAAATAGCTCCTCAAGAAGCACTATCAACG
GAATCAACTTGCCCTATAAACCAGTCATCTCATCAGC
TCTTCTCTTTCCAGAGATAAGTGGCAGCAAATTGAAC
TTTGAAGGCATTTTTTTTGGAAGTCAGTTATTTGATGT
AGTAACCTTAAAATGTTTGAGGAACATGGCACAGTTG
ATAGAAGTCAAGACTTGGGGTCCAAAAGATCTGAGTT
TAAATCCCTGCTCTGACCCTTAGGGGCTGTGTGACTA
CTCAACTTCTGCTAAGGTTTACCTGCCCAGTTACATAT
TACATTTGCATGGGTAAAGGGAATTCCCCTGCCAGTG
ATACTGCATATTCTTGATGTATTACTGTAACTCTATAT
TGTATCCTAATGTCTCCACTCTCCAATTATGAGGCTAT
TACAATCAGTTGTTGTCCTTTGTTTTGGAAGAGGACC
AAAATGGCATCACTATGTTGGGGTCAACTGTGTCTGA
CTGGCTGATCAGACCAATATGAGCTTGGAACATTCTA
CCCCAGAACGGGAGCAAATAATCCATGTGAACATCT
AGGGTAGAGATGTCTCTCAATGTGCCCATCTCATATT
TCCCCTACTTTCATGGAAAGAGCACTAGGCTAGAATT
CTAATCCCAGCTTAGCTGGCCACGGACTTAATCTCTG
TCTTTGACCGGATCACTTTGCTCCTCAGTTTCCTTCAC
TATGGAATGATCAGTTGGGATCAGGACAGGGGTAGG
GAACCTGTAGCCTTGAGACCACGTGGCCTCTAGGTCT
TCAAGTGCGACCCTTTGACTGAATCCAAATTTCACAG
TCCAAACCCCCTTCATAAAAGGATTTGTTCTGTATAA
CTTGGACTCAGTCAAAAAGCCGCACCCAAGGACCCA
GAAGGCCACATGTGACCTCAGGATCACAGGTTCCCCA
CCCCTGAACTAAGACATCTTTGAGGTCCCTTAACACT
CCAGTATTCTTGGTAGGGTTCTTTGTATGTGATATTTG
CTCAAGAGTACACGTTTGCTCTTAGGGTTACGAGATA
CGCATGTATGACAGT

The Below graphs show the analysis made on the data one of them shows the relation between the gene symbols and number of indirectly associated genes derived and the other shows the relation between number of abstracts taken and number of indirectly associated genes derived. The Fig. 5 shows the bar graph in which the number of abstracts are taken on x-axis and number of indirect associations is taken on y-axis and the blue lines depict the number of indirect associations for every sample of abstracts. The Fig. 6 shows the bar graph in which gene symbols are taken on x-axis and number of indirect associations derived for every gene symbol is taken on y-axis and the blue lines depict the number of indirect associations for every gene symbol.

TGGCCTCTAGGTCTTCAAGTGCGACCCTTTGACTGAATCCAAATTTCACAGTCCAAACCCCCTTCA
TAAAAGGATTTGTTCTGTATAACTTGGACTCAGTCAAAAAGCCGCACCCAAGGACCCAGAAGGC
CACATGTGACCTCAGGATCACAGGTTCC-CCACCCCTGAACTAAGACAT
TGGCCTCAAGGTCCTGAAGTTCCACCCTTTGACTGAATCCAAACTTCACTGAACAAATCCCCTTA
ATAAAAAGATTTGTTCCATAAAAGTTGGACTCAAATAAAATGCTGCACCCAAGGACCTAGAAGG
CTATATGTGGCCTCAAGGTAGCAGGTTCCTCCACCCCTGATAT-AGACAT

CAGGGGTAGGGAACCTGTAGCCTTGAGACCACGT---GGCCTCTAGGTCTTCAAGTGCGACCCTTT
GACTGAATCCAAATTTCACAG------TCC--------AAACCCCCTTCATAAAAGGATTTGTTCTGTATAAC
TTGGACTCAGTCAAAAAGCCGCACCCAAGGACCCAGAAGGCCACATGTGACCTCAGGATCACAG
GTTCCCCACCCCTGAACTA
CAGGGGTAGGGAGCCTGAGGCCTCGAGGCCACATACAGCCCTCTAGGTCCTCAAATACAGCCCTT
TGGCTGAATCCAAACTTCCCTCCAAATTTCCCAGAAAAAAAATCCCCTTAATGAAAGGATTTGTT
CTGTCAAACTTGGACTCAGTCAAAAAGTCGCACTTAAGGACCTAGAGCGCTACATGTGGCCTGGA
GGCCACAAGTTCCCCATCCCTGATCTA

Figure 4. Sample of the Indirectly associated gene sequences

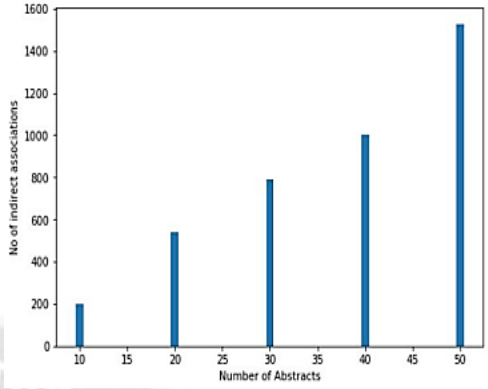

Figure 5 Bar graph between number of abstracts and indirect associations

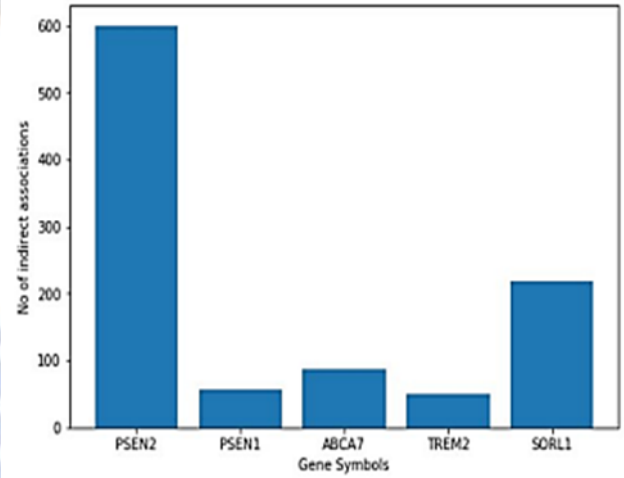

Figure 6. Bar graph between gene symbols and number of indirect associations

V. CONCLUSION

This paper introduces an approach to improve the identification of gene-disease associations crucial for biomedical research and drug development. Through biomedical text mining and frequent itemset mining, our method efficiently extracts disease-associated gene sets, prioritizing genes based on frequency counts and itemset sizes to enhance precision over existing techniques. Utilizing the NCBI database and Blast, indirect gene connections, especially those with single nucleotide polymorphisms, are established, extracting and processing gene sequences in XML format using NCBIXML. This methodology aims to uncover associations potentially missed by databases like HuGE Navigator, addressing current limitations. By employing frequent itemset mining, it enhances the accuracy of disease-gene extraction, unveiling novel relationships absent in mainstream databases. Emphasizing the need for advanced techniques in determining gene-disease correlations, this work underscores the potential for discovering precise, novel associations crucial for genetics research and targeted drug development. Future directions include exploring associations for more diseases, potentially utilizing evolving technologies like optical neural networks, and leveraging improved bioinformatics for discovering indirect associations.

REFERENCES

[1] S. Pletscher-Frankild, A. Pallej`a, K. Tsafou, J. X. Binder, and L. J. Jensen, "Diseases:Text mining and data integration of disease–gene associations," Methods, vol. 74, pp.83–89, 2015.

[2] J. Pi˜nero, N. Queralt-Rosinach, `A. Bravo, J. Deu-Pons, A. Bauer Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: a discovery platform for the dynamical human diseases and their genes," Database, vol. 2015, p. bav028, 2015.

[3] Sreekala S, K A Adbul Nazeer "A Literature Search Tool for Identifying Disease-associated Genes using Hidden Markov Model", 2014 First International Conference on Computational Systems and Communications (ICCSC).2015

[4] Wen-Juan Hou, Li-Che Chen, Chieh-Shiang Lu "Identifying Gene-DiseaseAssociations Using Word Proximity and Similarity of Gene Ontology Terms", 4th InternationalConference on Biomedical Engineering and Informatics (BMEI).2011

[5] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li "Networkbased globalinference of human disease genes" Molecular systems biology, 4(1).2008

[6] Jae-Yoon Jung, Todd F DeLuca, Tristan H Nelson, Dennis P Wall, "A literature search tool for intelligent extraction of disease-associated genes," Journal of the American Medical Informatics Association, Volume 21, Issue 3,2014 , Pages 399–405, https://doi.org/10.1136/amiajnl-2012-001563

[7] X. Wang, Y. Gong, J. Yi and W. Zhang, "Predicting gene-disease associations from the heterogeneous network using graph embedding," IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 504-511, doi: 10.1109/BIBM47256.2019.8983134.

[8] Opap K, Mulder N. "Recent advances in predicting gene-disease associations.".;6:578.2017. doi: 10.12688/f1000research.10788.1. PMID: 28529714; PMCID: PMC5414807.

[9] M. Sikandar et al., "Analysis for Disease Gene Association Using Machine Learning," in IEEE Access, vol. 8, 2020,pp. 160616-160626, doi: 10.1109/ACCESS.2020.3020592.

[10] Bhasuran B, Natarajan J. "Automatic extraction of gene-disease associations from literature using joint ensemble learning". PLoS One13(7):e0200699. 2018. doi: 10.1371/journal.pone.0200699. PMID: 30048465; PMCID: PMC6061985.

[11] Bravo, À., Piñero, J., Queralt-Rosinach, N. et al. "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research".BMC Bioinformatics 16, 55 . 2015. https://doi.org/10.1186/s12859-015-0472-9.

[12] Yang, H., Ding, Y., Tang, J. et al. "Identifying potential association on gene-disease network via dual hypergraph regularized least squares". BMC Genomics 22, 605(2021). https://doi.org/10.1186/s12864-021-07864-z

[13] [13] Liang H, Cao L, Gao Y, Luo H, Meng X, Wang Y, Li J, Liu W. "Research on Frequent Itemset Mining of Imaging Genetics GWAS in Alzheimer's Disease. Genes (Basel)". (2022) Jan 19;13(2):176. doi: 10.3390/genes13020176. PMID: 35205221; PMCID: PMC8871801.

[14] [14] Mutalib, Sofianita & Mohamed, Azlinah & Rahman, Shuzlina. "A Study on Frequent Itemset Mining for Identifying Associated Multiple SNPs". Journal of Computer Science & Computational Mathematics. (2019). 1-6. 10.20967/jcscm.2019.01.001.