ISSN 2347 - 3983
Volume 10. No.3, March 2022
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter031032022.pdf
https://doi.org/10.30534/ijeter/2022/031032022

GeneArticleAnalyser: A crucial part of the genome and word data analysis identification from Pub Med articles

Venu Paritala[1*], Harsha Thummala[2]

[1*]Agri It Solutions(Startup), Kovela Kuntla, Kurnool, Andhra Pradesh, India. vvenuparitala@gmail.com
[2]Dept.of Bbiotechnology, Vignan's Foundation for Science, Technology & Research (Deemed to be university)
hthummala99@gmail.com

Received Date February 06, 2022 Accepted Date : February 27, 2022 Published Date : March 07, 2022

ABSTRACT

A crucial parting of genome data analysis is identification from articles. Despite the availability of several articles in NCBI, researchers need to list out the genes, words, and quality of the genes from different articles, this process is highly time-consuming and requires a high throughput computing knowledge with advanced infrastructure. Hence, a user-friendly, single portal that could solve these shortcomings is a high prerequisite for genes & words analysis. Herein we introduce our indigenously developed web application, gene-article analyzer, for the effective analysis of abstract text file format/PubMed file format. It encompasses several innovative features to perform genes, gene quality analysis, top genes, words, word quality analysis, top words, and network of literature evidence. To our knowledge, the gene-article analyzer is the first of its kind to perform all these analyses in a single click. Gene-article analyzer proves to be a powerful tool for data analysis pipelines that are applied from various applications like big data analysis. Innovation in recent years has promoted marked progress in understanding genes. This review presents the analysis of biological data, scrutinizing approaches, and tools that tin give biological meaning to the data produced.

Availability and Implementation:

The gene-Article analyzer is available at:

https://venuparitala.shinyapps.io/gene-articleanalser/

Supplementary information*: Supplementary data is available at Bioinformatics online.*

Key words*:* Genes, Analysis, Tool, Network, Pub med

1.INTRODUCTION

Gene is the functional unit of heredity; it is made up of DNA. It is present in the cell nucleus; each person has two copies of genes come a single copy from each parent. Some genes give instructions to make molecules called proteins. In humans([1], [2]), genes may vary in size from person to person from a couple of hundreds to more than 2 million DNA bases. Most of the genes are the same in almost all people, in gene expression, DNA is copied first into RNA, and then RNA tin is directly functional. Genes are organized in one after the another/linear order, in structures called chromosomes[3], complete set of genes of a cell or organism called the genome, the coding sequence region of a gene is known as the coding sequence, which is the parting of a gene's DNA or RNA that codes for a protein[4]. To date, a large number of bioinformatics tools are available for analyzing data[5].

Bioinformatics techniques have been developed to determine true variants and weed out false positives and negatives[6,7]. Scientists will benefit from these developments promptly, requires bio-informatics software to match with new requirements[8,9]. We developed a unique tool for the prediction of genes presents in an uploaded abstract/file as well as what are the most used genes by researchers and the quality of the genes([10], [11]). This software builds to make it researchers' life easy to collect the data in a shorter full stop of time[12]. To gain these insights, researchers do not require high-throughput computing infrastructure, knowledge of various computational languages, and bioinformatics tools([13], [14]). Hence, a user-friendly, click-based, single portal is a highly prerequisite for variant

analysis[15]. Recently the emergence of web-based bio-informatics applications has proved to be promising in overcoming barriers such as computational infrastructure, platform, and programming language dependencies[16].

2.MATERIALS AND METHODS

2.1 Implementation

Gene- Article analyzer web application is developed with R (Reproducible Research with R and RStudio, Second Edition, 2018). The server is hosted in a cloud platform. For the development of the analytical pipeline, various R packages were used. The packages are shiny, Shiny themes, Shinycssloaders, Shiny custom loader,DT, ggplot2, Pubmed. MineRggnet work, igraph, shiny custom loader, HTML tools, ggrepel,html widgets. It contains both the user interface and server[17].

Shiny: It is helpful to create a UI and server for applications/databases, **Shiny themes**: themes used with shiny to beautify the application, **Shinycssloaders**: adds loading animations, **Shinycustomloader**: A custom CSS/HTML or gif/image file for the loading screen in R 'shiny', **PubMed.mineR:** Used to mine the data from Pubmed article, **ggplot2**: It plots graphs to given data, **DT**: Description Data objects in R can be rendered as HTML tables using the. JavaScript library Data Tables'(typically via R Markdown or Shiny)([22], [23]),**ggnet work**: ggnetwork package, which provides several geoms to plot network objects with ggplot2, **igraph**: graph is a library collection for creating and manipulating graphs and analyzing networks, **ggrepel**: ggrepel provides geoms for ggplot2 to repel overlapping text labels, **shinymanager**:Simple and secure authentication mechanism for single Shiny applications[18],**htmltools**: Tools for html generation and output in R, **htmlwidgets**: A framework for creating HTML widgets that render in various contexts including the R console.

2.2 Input

Gene- Article analyzer takes of Abstract text file format/PubMed file format for processing as input. The preferred Abstract text file format/PubMed file format version is 4.3; any other version or file type will result in an error message([19], [20]). A provision to browse and upload a file from the local computer is provided for the convenience of the user. The maximum file size that can be uploaded by the user is <5Gb. Upon successful valid PubMed file upload, Gene- Article analyzer reactively performs quality checks. All other analysis is performed by clicking the corresponding tabs([24], [25]). The major components that are included in the Gene-Article analyzer **are** as follows 1. genes, 2. Gene quality check, 3. top genes, 4. words, 5. words quality check,6. Top words, 7. network, 8. about. User can download results for each analysis as JPEG/CSV formats. To demonstrate the Gene-Article analyzer, we provided an example file in the webserver. Users can download and explore each component using this file[26].

2.3 Enrichment

The enrichment components are explained as follows it seen Figure 1.

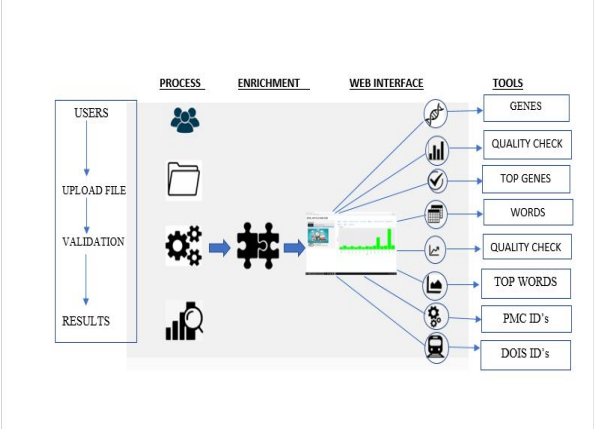

Figure 1: Flow chart of Gene-Article Analyser tool components.

3.RESULTS

3.1 Genes

It is a vital process in NCBI articles required to identify gene and their biological, biochemical functions[21]. This component performs the list of genes of a given Abstract text file/PubMed file. The results obtained from the gene list tool can be downloaded in CSV format to explore the reference gene frequency, gene names, and gene symbols. Further, the results are displayed as an interactive (seen Table 1) and a search box on the top left corner of the gene table provides search queriesof user interest. Ex: Covid-19

Table 1: Genes of Covid-19 disease

Gene_symbol	Genes	Freq
ACE2	angiotensin I converting enzyme (peptidyl-dipeptidase A) 2	770
SARS	seryl-tRNA synthetase	415
T	T, brachyury homolog (mouse)	307
CRP	C-reactive protein, pentraxin-related	144
NHS	Nance-Horan	117
TMPRSS2	transmembrane protease, serine 2	89
ACE	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	52
NPS	neuropeptide S	30
CD4	CD4 molecule	26

3.2 Genes Quality analysis

This component performs the quality check plot of a given Abstract text file/PubMed file. The results obtained from the quality check (seen Figure 2)tool can be downloaded right click on the image to save png/jpg format to explore the reference gene frequency and gene symbol. Ex: Covid-19

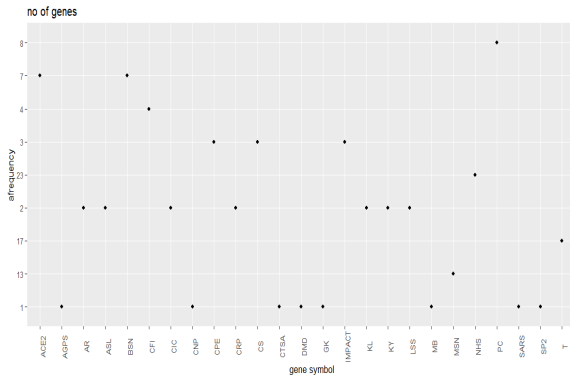

Figure 2: Genes quality factor of Covid-19

3.3 Network

This tab panel contains the network connection between the genes data.In network Yellow Node are genes and blue lines are their pathway. It observe seen Figure 3 Ex: Covid-19

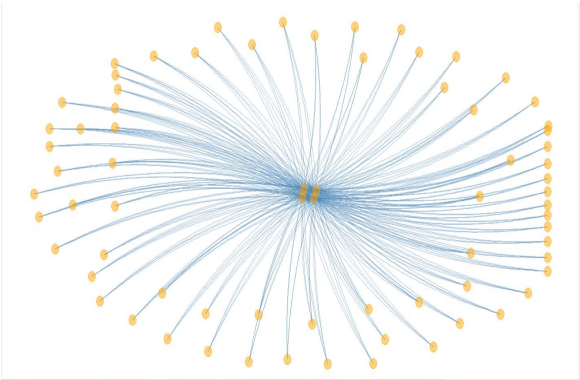

Figure 3: Gene network of covid-19

3.4 Top genes

This component performs the top genes plot of a given Abstract text file/PubMed file. The results obtained from the top genes plot tool can be downloaded right click on the image to save png/jpg format. To explore the reference gene frequency and gene symbol. These plots mainly explore based on the gene's frequency. These tab panels perform two (seen Figure 4) one is top 20 genes and another one is top 30 genes. Ex: Covid-19

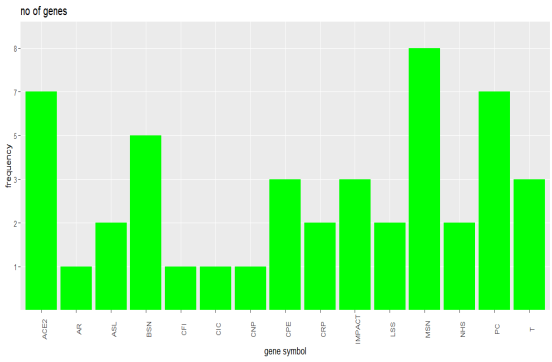

Figure 4: Top genes of covid -19 disease

3.5 Words

These tools mainly work when users upload an article to count how many words to repeated to list out them. This tool lists out the words, word frequency. This helps the user to effortlessly find the gene of interest in the entire Abstract text file/PubMed file. A search box is provided to identify the word of interest thereby minimizing the time. The

results (seen Table 2) can be downloaded in . CSV format for further exploration. Ex: Covid-19

Table 2: Words data of covid-19 disease

SI.NO	words	Frequency
1	covid-19	68
2	2020	45
3	disease	45
4	skin	43
5	university	18
6	pandemic	15
7	coronavirus	15
8	DOI	15
9	medical	15
10	medicine	13
11	health	131
12	information	13
13	patients	13
14	sars-cov-2	12
15	china	12
16	respiratory	10
17	treatment	10

3.6 Word Quality analysis

This component performs the quality check plot of a given Abstract text file/PubMed file. The results obtained from the quality check (seen Figure 5) can be downloaded right click on the image to save png/jpg format to explore the reference word frequency and word name. Ex: Covid-19

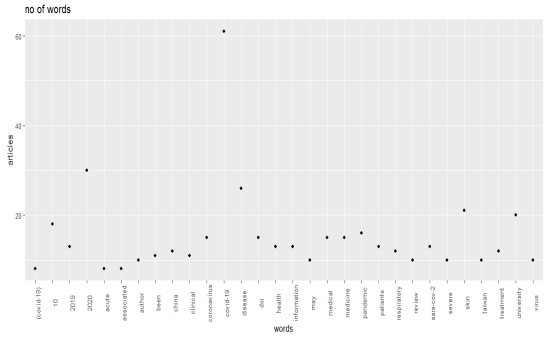

Figure 5: Words qulity factor of covid-19

3.7 Top Words

This component performs the top genes plot of a given Abstract text file/PubMed file. The results obtained from the top genes plot tool can be downloaded right click on the image to save png/jpg format. To explore the reference word frequency and word symbol. These Figure 6 mainly explore based on the word frequency. These tab panels to perform to plot one top 20 words and another is 30 words. Ex: Covid-19

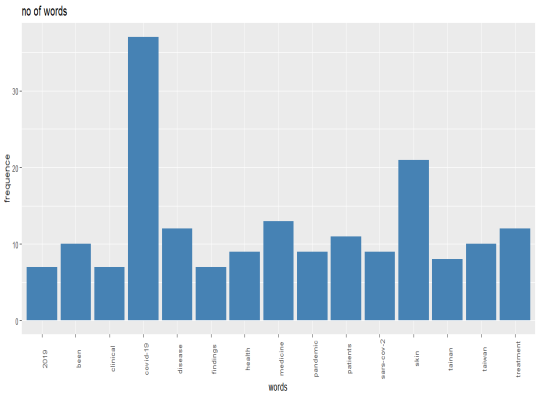

Figure 6: Top words of covid-19 disease

3.7 PMC ID

This component performs when a user uploads an abstract file, it gives PMCIDs for each abstract. It Observe in Table 3.

Table 3: PMC ID of Covid-19 disease abstract

SI.NO	PMC ID
1	PMCID: PMC7361342
2	PMCID: PMC7775718
3	PMCID: PMC7235502
4	PMCID: PMC7267323
5	PMCID: PMC7795815
6	PMCID: PMC7386392
7	PMCID: PMC7536131

3.8 DOIS ID

Digital Object Identifier ID is a component that can give DOIS ID for each abstract when a user uploads

an abstract text file. It sample results dispatches in Table 4. Ex: Covid-19

Table 4:DOIs ID of covid-19 abstracts

SI.NO	DOIs ID
1	DOI: 10.1002/JMV.26232
2	DOI: 10.3906/SAG-2005-182
3	DOI: 10.1111/DTH.13430
4	DOI: 10.1002/JMV.25965
5	DOI: 10.1080/14787210.2020.1797487
6	DOI: 10.3390/MOLECULES26010039
7	DOI: 10.1016/J.DISAMONTH.2020.101058
8	DOI: 10.1136/POSTGRADMEDJ-2020-138386
9	DOI: 10.1016/J.AMJMS.2020.10.002

4.CONCLUSION

Gene- Article analyzer (seen Figure 7) can be used to analyze and visualize genes and words. Further, provide a significant amount of information to interpret validate high-performing genetic studies. Case study example, database benchmarking, and comparisons to existing software are available in Supplementary material 1. Later this tool will be expanded for the processing of other model organisms.

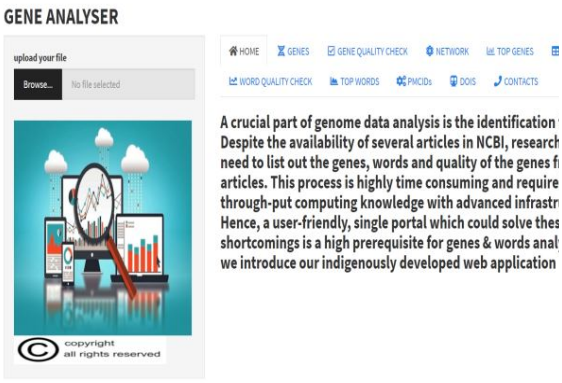

Figure 7: The interface of Gene-Article Analyser Tool.

ACKNOWLEDGMENT

Author thanks DST FIST facility provided by the Vignan's institute for science, technology, and research for the execution of work,

REFERENCES

1. Krwawicz J, Arczewska KD, Speina E, Maciejewska A, Grzesiuk E. Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease.Acta Biochimica Polonica. 2007;54(3):413–434. [PubMed] [Google Scholar]
2. de Bont R, van Larebeke N. Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis. 2004;19(3):169–185. [PubMed] [Google Scholar]
3. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996, 17:37–54.
4. Smyth P. Data mining: Data analysis on a grand scale? Stat Methods Med Res 2000, 9:309–327.
5. Lovell MC. Data mining. Rev Econ Stat 1983, 65:1– 11.
6. Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann; 2006.
7. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2008.
8. Engelbrecht AP. Computational Intelligence - An Introduction. Chichester: John Wiley; 2007.
9. Vesset D, McDonough B. Worldwide business intelligence tools 2008 vendor shares, IDC Competitive Analysis Report (2009).
10. Frank E, Hall M, Holmes G, Kirkby R, Pfahringer B, Witten I. Weka: A machine learning workbench for data mining. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. New York: Springer; 2005, 1305–1314.

11. Goebel M. A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations. Newsletter 1999, 1:20–33.
12. Wang J, Hu X, Hollister K, Zhu D. A comparison and scenario analysis of leading data mining software. Int J Knowl Manage 2008, 4:17–34.
13. P. Agarwal and V. Bafna "The ribosome scanning model for translation initiation: Implications for gene prediction and full-length cDNA detection," Intelligent Systems for Molecular Biology 6, 2–7 (1998).
14. V. B. Bajic, S. Tang, H. Han, V. Brusic and A. G. Hatzigeorgiou "Artificial neural networks based systems for recognition of genomic signals and regions: A review," Informatica 26 389–400 (2002).
15. P. Baldi, S. Brunak, Y. Chauvin and A. Krogh, "Hidden Markov models for human genes: Periodic patterns in exon sequences," in Theoretical and Computational Methods in Genome Research, pp. 15–32 (1997).
16. P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes, Bioinformatics 17, 509–519 (2001).
17. M. S. Boguski, T. M. J. Lowe, and C. M. Tolstoshev, "dbEST — database for 'expressed sequence tags," Nat. Genet. 4, 332–333 (1993).
18. M. S. Boguski and C. M. Tolstoshev, "Gene discovery in dbEST," Science 265, 1993–1994 (1994).
19. M. Borodovsky and J. D. McIninch, "GENEMARK: Parallel gene recognition for both DNA strands," Computers and Chemistry 17(2), 123–133 (1993).
20. M. Borodovsky, J. D. McIninch, E. V. Koonin, K. E. Rudd, C. Medigue and A. Danchin, "Detection of new genes in a bacterial genome using Markov models for three gene classes," Nucleic Acids Res. 23, 3554–3562 (1995).
21. L. Breiman, L. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Wadsworth and Brooks, 1984.
22. R. J. Brooker, Genetics: Analysis and Principles. Addison-Wesley, Reading, MA, 1999.
23. J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, CA, 2000.
24. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, Berlin, 2001.
25. D. Heckerman, "Bayesian networks for knowledge discovery," in Advances in Knowledge Discovery and Data Mining, pp. 273–305. MIT Press, Cambridge, MA, 1996.
26. D. Hennessy, B. Buchanan, D. Subramanian, P. A. Wilkosz and J. M. Rosenberg, "Statistical methods for the objective design of screening procedures for macromolecular crystallization," Acta Crystallogr. D Biol. Crystallogr. 56, 817–827 (2000).