Quest Journals
Journal of Software Engineering and Simulation
Volume 8 ~ Issue 8 (2022) pp: 07-14
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Python for Ranking Hub Proteins

S I Aruna[1*], S Sujatha[2], and E S Neenu[3]

1(Research Scholar, Registration No:19213082022001, Interdisciplinary Research Centre, Department of Biotechnology, Malankara Catholic College, Mariagiri, Affiliated to ManonmaniamSundarnar University, Tirunelveli – 627012, Tamilnadu, India)*
2. (Assistant Professor, Department of Biotechnology, Malankara Catholic College, Mariagiri)
3. (Programmer, Indriyam Biologics Pvt. Ltd. SCTIMST – TIMed, 5th floor, BMT wing, Poojappura, Thiruvananthapuram.)

ABSTRACT : *In this paper,Hub proteins are proteins that interact with numerous other proteins. Because of their tremendous interconnectedness, hub proteins are crucial in protein-protein interaction. However, identifying these hub proteins based on their function is a difficult undertaking. The Markov Clustering Algorithm is often used to rank the hub proteins in investigations. The hub proteins are ranked according to how connected they are. The goal of this study is to rank hub proteins according to their interacting partners because as the number of interacting proteins rises, so does their biological significance. If we rank the hub proteins according to their interacting partners, we can quickly identify the hub protein that is biologically significant. Pandas, a free and open-source python library, is used to rank the hub proteins. The ranked hub proteins are examined and categorised as disease-causing or not, heat shock or cold shock, utilising pandas and tkinter. Hub proteins are responsible for complicated ailments, the most significant of which is neurodegenerative disease. Hence a database called Neuro Hub was constructed using javascript, CSS, and HTML so that we could examine the hub proteins responsible for neurodegenerative diseases and related information. It will make it easier for people to get data on this topic and will also provide more details for those doing research on neurodegenerative diseases.*

KEYWORDS: *Hub proteins, Python, Pandas, Tkinter, Neuro Degenerative Diseases, Javascript*

R**eceived 07 August, 2022; **R**evised 20 August, 2022; **A**ccepted 22 August, 2022 © The author(s) 2022. **P**ublished with open access at **www.questjurnals.org

I. INTRODUCTION

A network consists of several nodes that are connected by edges. Hubs are the name for these nodes [1]. For instance, proteins act as nodes and interactions as edges in a protein-protein interaction. Small, densely connected nodes make up a scale-free network [2]. The key benefit of this is the domain repeats which are associated with binding are enriched in hubs. The hubs are divided into two groups—party hubs and dating hubs—based on the expression profile. The party hubs engage with the majority of partners simultaneously [3]. Party hubs are the central or static part. Party hubs are the central or static part. They are long disordered region compared to date hubs; indicating that the regions are crucial for flexible binding. They serve as the central node of densely grouped functional modules. However, because the date hubs are dynamic, they cannot communicate with the majority of partners at same time [4]. In every sector, hub proteins are very important. Hubs are particularly intriguing pharmacological targets since they are important for cancer research and have unique biological characteristics that make them more important than non-hub proteins. Hub proteins play a crucial part in the modular structure of the protein interaction network. Hubs have fast turnover and regulation despite being slow to evolve because of evolutionary conservation. They play a fundamental part in a wide range of biological processes in a number of different ways, and they are also to responsible for a number of diseases like cancer, auto immune disorders, and neuro degenerative illnesses [5]. However, identifying these hub proteins based on their significance is a difficult undertaking.

The relevance of hub proteins nowadays is extremely high, making it challenging to rank them in terms of importance. Currently there is a database STRING, in that the hubs are ranked on the basis of their interaction [6]. The Markov Clustering Algorithm (MCL) is an algorithm that may be used to rank the hubs based on the

characteristics of their shared connectivity [7]. Although the MCL algorithm can be used in low-density locations, hubs can still be located there. For exploring significant nodes in the interactome network, there is another web-based service called Hub object analyzerHubba [8], which is based on the Maximum Neighbourhood Component method (based on topology). However, it is extremely particular for disease pathologies and depends on the correctness and completeness of the supplied interactome information. Although ranking hubs is now very tough, they are very significant.

Through this work, we intend to rank the hub proteins according to the proteins that they interact with; as the number of interacting proteins rises, so will the hub proteins' biological significance. Numerous studies have shown that the biological significance of the interacting proteins depends on the biological functions of the hub proteins; the more interacting proteins, the more biological relevance. Python's major strength is its enormous library collection and open-source nature, which make it a general-purpose language [9]. Pandas and tkinter (a Python GUI application), out of all the libraries, were utilized to implement the work. Using the _.sort value() method and the panda library in Python, the interacting proteins with the highest value are identified. The above-ranked hub proteins' susceptibility to heat shock or cold shock is also determined using the Python computer language.

Numerous complicated diseases, including cancer, autoimmune disorders, and neurodegenerative disorders, are caused by the hub proteins [10]. Since neurodegenerative disease is the most prevalent of these, a database for it was also created using JavaScript, CSS, and HTML. Using this database, we can examine the hub proteins responsible for neurodegenerative diseases, different neurodegenerative disorders, their symptoms, pathways, and treatment options, among other things.

II. MATERIALS AND METHODS

2.1 Python

Guido Rossum founded the open-source general-purpose language Python in 1989. Python supports procedural and object-oriented programming. It is perfectly suited for the quick prototyping of sophisticated applications. The most recent version of the Python language, Python 3.x, is widely used in the software industry. Python language usage was prevalent in many large software organisations, and it appears that this trend is continuing. Various businesses, including Google, YouTube, Amazon, Facebook, Instagram, Uber, Dropbox, etc. The Python programming language is utilised in a variety of fields, including bioinformatics and machine learning applications. One of Python's greatest strengths is its extensive library system. Pandas and Tkinter were two of the libraries we used to implement the work. Over the years, many versions of Python have been released; the most recent is version 3.10.4.[9]

2.1.1 Pandas

Wes McKinney created the open-source Pandas Python library in 2008 because he was in need of a powerful, adaptable tool for data analysis. Pandas is primarily used to deal quickly and intuitively with labelled data as well as relational data. Since Pandas delivers high-performance data manipulation and analysis tools using its potent data structures, it is frequently used for data science, data analysis, and machine learning. The NumPy library, which supports multi-dimensional arrays, is the foundation upon which this library is based. Pandas and Python are widely used in both academic and industrial fields like finance, economics, statistics, analytics, etc. Pandas have been released in a variety of versions over the years. The latest version of the pandas is 1.4.1. [9]

Pandas uses the two data structures series and data frame to manipulate and analyse data. One-dimensional labelled arrays using the Pandas series can carry any datatype, including integers, strings, floats, and Python objects, among others. The term indexes refers to the axis as a whole. An excel sheet has a column called "Pandas Series." The two-dimensional data structure, where data is arranged in rows and columns in a tabular form, is implemented using a Pandas DataFrame. Data frame id size-mutable, potentially heterogeneous, and having marked axes, DataFrame is a tabular data format (rows and columns). There are three main parts to it: the data, rows, and columns. The datasets will be loaded from pre-existing storage, such as an Excel file, SQL database, or CSV file, to build both Pandas Series and Pandas Data Frame. It can be generated from scalar values, lists, dictionaries, etc [11].

2.1.2 Tkinter

One of the most widely used Python GUI (Graphical User Interface) packages is Tkinter, which is used to create desktop programmes. The Python interface for Tk, the GUI toolkit for Tcl/Tk, is called Tkinter. It combines the common GUI frameworks of Python and Tk. The quickest and easiest approach to construct GUI apps is with Python and Tkinter. To create GUI programmes, several different programming languages use the open-source, cross-platform Tk toolkit. Python has long Tk/Tcl as a core component. It offers a powerful and

platform-independent windowing toolkit that comes with the default Python library pre-installed. Tkinter has seen the release of numerous versions over time. The latest version of the tkinter is 8.6 [12].

2.2 HTML

Web pages are made using the Hyper Text Mark – up Language. It is a markup language, not a programming language. A set of markup tags are used in markup languages. To describe web pages, HTML uses markup tags. HTML has the ability to contain embedded programming language code (like JavaScript) that can change how Web browsers and other processors behave. A default or homepage is given to the user in HTML format when a web client accesses a website. The HTML file can show multimedia objects like pictures, music, and videos. The HTML document does not actually include the objects. Instead, a text reference to a photo or other multimedia object is added externally [13] to the HTML document.

2.3 JAVASCRIPT

JavaScript is a scripting language used to enable programmatic access to objects within other applications. It is primarily used in the form of user JavaScript for the development of dynamic websites. JavaScript is characterized as a dynamic, weakly typed, prototype based language with first class functions. JavaScript is used to do user side validation. It is used to create mouse over events [13].

2.4 CSS

Cascading Style Sheets also used provide more visual impact to website. It helps to add design layer for the HTML program. It is a styling method, used to add more style and structure to web site [13, 14].

III. METHODOLOGY

For the programming, Python language is used version Python 3.7.9 with the operating system Windows 10 (64-bit OS, x64-based processor). Python library such as pandas 1.3.5 and tkinter 8.6 was used and Notepad ++ was used as editor.

3.1 To find the higher interacting proteins

The hub protein with the highest values of interacting proteins is found using the Python programming language. Using the _.sort value() method and the panda library in Python, the interacting proteins with the highest value are identified. A dataset concerning hub proteins is contained in an excel file that is provided as an input file. In the python code, the _.read excel() function is used to read the excel file, and the result is stored into another excel file using the _.ExcelWriter() function, which writes the data into excel file. Python's tkinter module is used to build a GUI for user interaction. An interface is designed to allow users to add an input file (an excel file) containing information on hub proteins, from which we wish to determine which interacting proteins have a higher value.

3.2 To check whether the hub protein causes disease or not

To determine whether or not the hub protein causes disease, Python computer language is employed. They gathered information on hub proteins and other hub protein details, which they then recorded in an excel file. This excel file is used as the input file to determine our goal. The software is configured such that a user can choose a specific hub protein to determine whether or not it contributes to a disease. The outcome is then displayed along with some fundamental facts about the protein. The input file, which is an excel file, serves as a dataset in this case, and the input hub protein is checked throughout it to determine the goal. Python code is constructed at the back end to determine whether the hub protein causes sickness or not, and a GUI is created using the Tkinter toolkit in Python. To successfully run the code, other libraries like pandas are utilised in conjunction with the tkinter library to retrieve and process the excel file.

3.3 Database creation

The suggested system was developed using HTML, CSS, and Javascript. The software application's output in this system is a database with details on neurodegenerative illnesses. A user can choose an illness and find information about it using a straightforward keyword search. According to the user's requirements, the software should extract the disease from the database. Before sending a query to the database, user input needs to be verified. The information that has been pulled from the database must be displayed as a webpage with pertinent information. Analysis of the system shows the need for database application, efficient server side programming, user friendly web interface, client and administrator side validation and database connectivity. The programming language Javascript complies with all coding standards. It features a lot of modules for working with strings, connecting to databases, programming on the server side, and programming on the

administrator side. HTML can be used to construct web interfaces. The server-side programme must validate the user-entered query before passing it as an HTTP request. JavaScript is compatible with HTML and can be used for client validation.

IV. RESULT AND DISCUSSION

Physical interactions between proteins are cardinal to biological processes. In order to restore their function, proteins must interact with one another. Hub proteins are densely interconnected proteins that have a wide range of biological significance and are also to responsible for a number of illnesses, including cancer, auto immune disorders, and neurodegenerative diseases. Hub proteins were ranked in early investigations based on connection and interaction. String is an already available database for hub proteins in that the hub proteins are organised according to how they interact with one another. The Markov Clustering Algorithm is an algorithm that aids in ranking the hubs based on the characteristics of their shared connectivity. The hubs are located in the denser zone, however the MCL algorithm is applicable for low-density regions. For investigating significant nodes in the interactome network, there is another web-based tool called Hubba that also uses MNC algorithm, which is typically based on topology. However, it depends on how accurate and complete the input interactome database is. We are unable to predict the precise outcome of any of the aforementioned methods because they all use graph-based analysis to interpret hub protein clusters. It is also impossible to interpret each protein's functions based on biological precedent. However, using Python, we can quickly order hub proteins according to functional relevance; all that is needed is the number of interacting partners.

Through this investigation, the hub proteins were gathered and ranked according to the number of interacting partners. The biological relevance will increase based on the interacting partners. Biological function will also rise as the number of interacting partners increases. The classification method developed with the use of Tkinter and Panda also aids in determining which family the listed hub proteins belong to, whether they are heat shock or cold shock proteins, if they cause any diseases or not, and also displays their molecular weight. It also makes it very simple to identify the disease-causing hub proteins. The key draw of this work is the use of simple python modules to do all of those tasks.

There are numerous databases on neurodegenerative diseases, however none of them provide information on hub protein-related neurodegenerative diseases. Hence a database called "Neuro Hubs" was also constructed because the hub proteins are also responsible for many complicated disorders like cancer, autoimmune diseases, and neurodegenerative diseases. Neuro Hub contains information on the hub proteins that cause neurodegenerative disorders, as well as the several types of neurodegenerative diseases, their causes, symptoms, pathways, and treatments. It will be informative for both the general audience and professionals working on neuro degenerative disease research and analysis.

4.1Working of the GUI to find the higher interacting proteins is given below:

Figure 1: Python code is executed using CMD (Command Prompt)

Figure 2: GUI will be open up to upload the input file with the data of hub proteins using the upload file button.

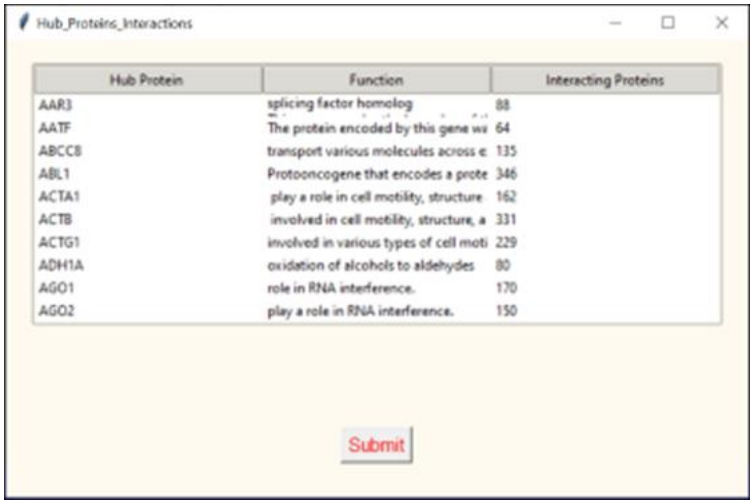

Figure 3: The input data is displayed in the interface and submit button selected to find the protein with higher interacting value.

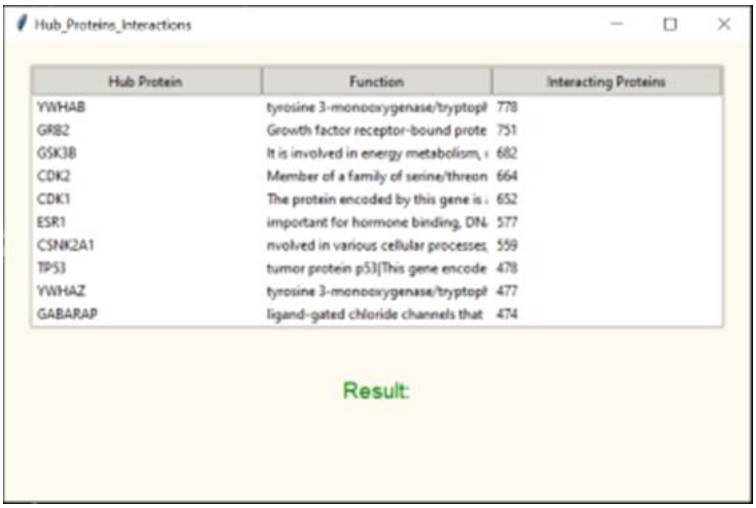

Figure 4: Finally, we will get the result of proteins with higher interacting values

4.2 Working of the GUI to check whether the hub protein causes disease or not:

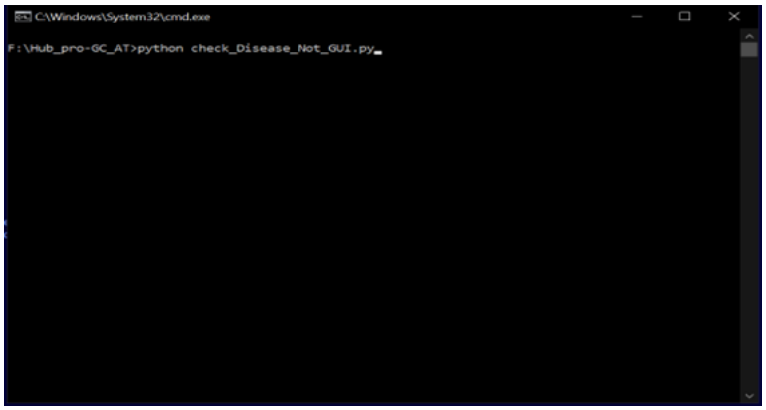

Figure 5: Python code is executed using CMD (Command Prompt)

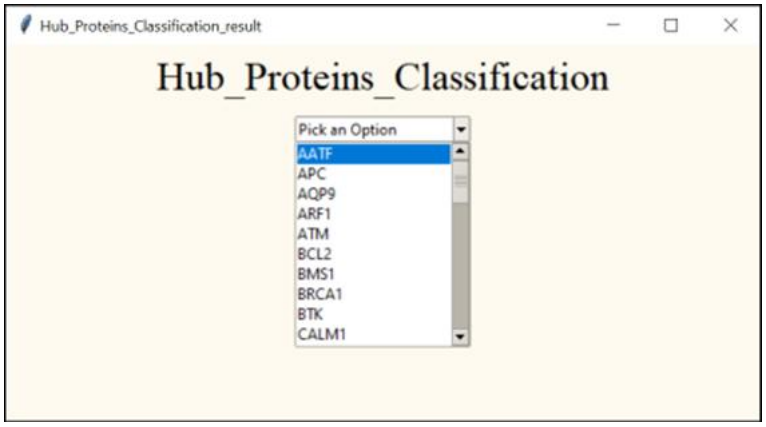

Figure 6: From the GUI interface got, select the hub protein that we need to check from the hub proteins is given. so that the user can select the protein to check.

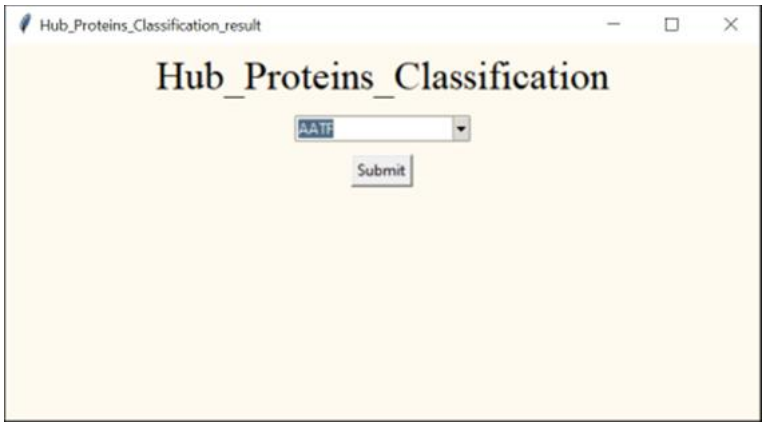

Figure 7: For example, from the options 'AATF' is selected.

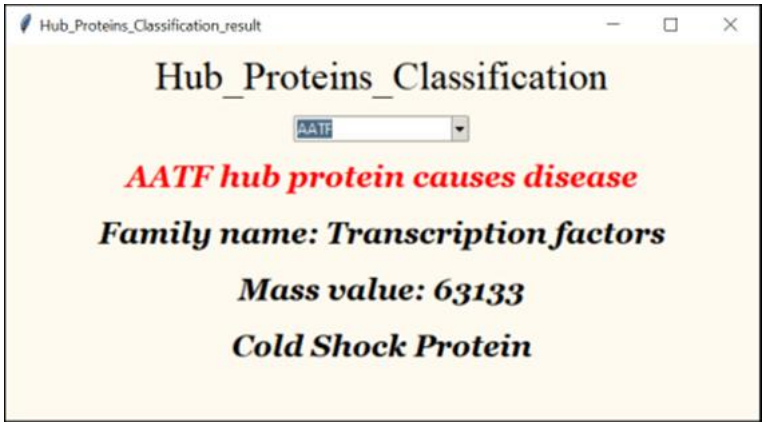

Figure 8: Output is displayed along with the information about the selected protein such family name, Mass value, and heat shock protein or cold shock protein.

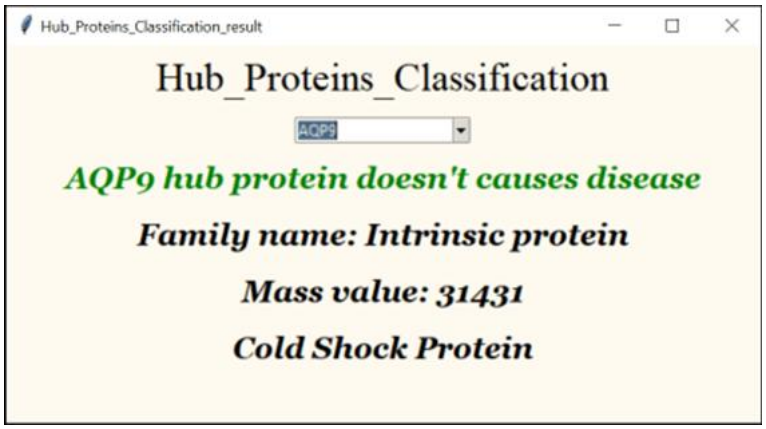

Figure 9: Another example, from the options 'AQP9' is selected Output is displayed along with the information about the selected protein such family name, Mass value, and heat shock protein or cold shock protein.

4.3Database creation:

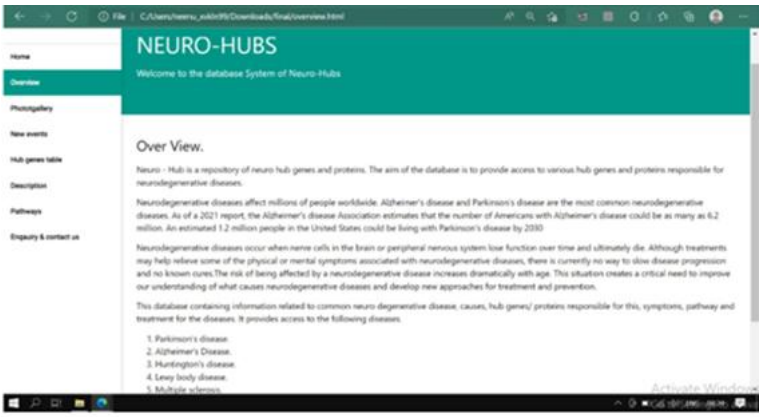

Figure 10: Home page of Neuro Hubs database

V. CONCLUSION

The biological significance grows with an increase in the number of interacting partners, and this work helped to rank the hub proteins according to their interacting partners. Prior to now, complicated techniques based on the connectivity and topology of proteins were used to rank the proteins. However, this study ranked the hub proteins on the basis of their interacting partners. Python, a straightforward programming language, was also used. With the help of Python and its modules, a classification system was also developed after ranking, allowing us to quickly determine whether the aforementioned hub protein causes disease or not and whether it is caused by heat shock or cold shock. Future researchers who want to learn more about hub proteins and their biological significance will find this study to be illuminating.

Some serious diseases, including cancer, autoimmune illnesses, and neurodegenerative disorders, are caused by hub proteins. This research also led to the creation of the "Neuro Hub" database. It contains comprehensive details about hub protein-related neurodegenerative diseases. It will make it easier for people to get data on this topic and will also provide more details for those doing research on neurodegenerative diseases.

REFERENCES

[1]. Ekman, D., Light, S., Björklund, Å.K. et al. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?. Genome Biol **7,** R45 (2006). https://doi.org/10.1186/gb-2006-7-6-r45

[2]. He, X., & Zhang, J. (2006). Why Do Hubs Tend to Be Essential in Protein Networks? PLoS Genetics, 2(6), e88. https://doi.org/10.1371/journal.pgen.0020088

[3]. Safari-Alighiarloo N, Taghizadeh M, Rezaei-Tavirani M, Goliaei B, Peyvandi AA. Protein-protein interaction networks (PPI) and complex diseases. Gastroenterol Hepatol Bed Bench. 2014 Winter;7(1):17-31. PMID: 25436094; PMCID: PMC4017556.

[4]. Kenley, E. C., Kirk, L., & Cho, Y.-R. (2011). Differentiating party and date hubs in protein interaction networks using semantic similarity measures. Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine - BCB '11, 641. https://doi.org/10.1145/2147805.2147916

[5]. Wolfson, M., Budovsky, A., Tacutu, R., & Fraifeld, V. (2009). The signaling hubs at the crossroad of longevity and age-related disease networks. The International Journal of Biochemistry & Cell Biology, 41(3), 516–520. https://doi.org/10.1016/j.biocel.2008.08.026

[6]. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 49(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074

[7]. Zaki, N., Berengueres, J., & Efimov, D. (2012). Detection of protein complexes using a protein ranking algorithm. Proteins: Structure, Function, and Bioinformatics, 80(10), 2459–2468. https://doi.org/10.1002/prot.24130

[8]. Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., & Lin, C.-Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Systems Biology, 8(S4), S11. https://doi.org/10.1186/1752-0509-8-S4-S11

[9]. Mark Lutz. Learning Python: Powerful Object-Oriented Programming (5th ed.). O'Reilly Media, Inc. 2013.

[10]. Senadheera, S. P. B. M., & Weerasinghe, A. R. (2020). Hub Genes Identification in Brain Cancer with Gene Expression Data. 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), 125–130. https://doi.org/10.1109/ICTer51097.2020.9325446

[11]. McKinney, W. & others, 2010. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference. pp. 51–56.

[12]. Bezerra Beniz, D and Espíndola, Alexey (2016). Using Tkinter of Python to create Graphical User Interfaces (GUI) for scripts in LNLS Architecture Overview Source Code Excerpts Tkinter Solution of DXAS. https://doi.org/10.13140/RG.2.2.14230.86084

[13]. Laura Lemay, Rafe Colburn, & Jennifer Kyrnin. Mastering HTML, CSS & JavaScript Web Publishing (1st ed.). BPB Publications; 2016

[14]. Wilson, D., Hassan, S.-U., Aljohani, N. R., Visvizi, A., & Nawaz, R. (2022). Demonstrating and negotiating the adoption of web design technologies: Cascading Style Sheets and the CSS Zen Garden. Internet Histories, 1–20. https://doi.org/10.1080/24701475.2022.2055274