Vol I December 2021
E-ISSN xxx xxx
pp. 14-20

Utilization of Graph Theory Modelling in Searching Potential Indonesian Medicine Plants as Antihypertension

Deden Ardiansyah [1, a)], Lusi Agus Setiani[2, b)], and Arif Rahman Hakim [3]

[1] *Computer Engineering, Vocational School Pakuan University, West Java, Indonesia*
[2, 3]*Pharmacetical Department, Faculty Of Mathematic and Science, Pakuan University West java Indonesia*

[a)] Corresponding author*: ardiansyahdeden@unpak.ac.id*

Abstract. Hypertension can be defined as a disease characterized by a systolic blood pressure value of at least 140 mmHg and or a diastolic blood pressure value of at least 90 mmHg. Broadly speaking, hypertension is divided into 2 types. Tissue Pharmacology, based on the concept of multidisciplinary integrity, is a powerful tool for analyzing multi-level tissue from molecular-target-pathways-of-disease through interactions between traditional medicine and disease from a holistic perspective The materials used in this study include: search for proteins related to hypertension at Uniprot, download results of interaction designs of compounds and proteins on STRING, cluster results from the Cytoscape application, download results of potential substance data on KNApSAcK, results of data downloads information on substances in PubChem, results of RStudio processing, results of data searches, extraction and labeling from Jupyter Lab, results of downloading of interaction designs of compounds and proteins on STITCH, which will later be analyzed and combined using the Orange application. The search process for significant proteins related to hypertension is carried out in the Uniprot database on the initial web display. There is a display of protein search results related to hypertension select on the left that reads "Homo sapiens" because the search for hypertension protein is only limited to proteins found in humans. The calculation of overall centrality is done by entering a script/command in the kernel in the Python program. The results obtained are that the AGTR1_HUMAN protein has an overall centrality value of (2.76967082423431), the GNAS_HUMAN protein has an overall centrality value of (-1,38194807757488), ACE2_HUMAN protein has an overall centrality value of (-1,72856432286345), TAC3_HUMAN protein has an overall centrality value (-1.01621749330442), EDN1_HUMAN protein has an overall centrality value of (-0.392822941980431). These values indicate how significant a protein is in a cluster. The results of the search for the best 1 plant found 5 plants that can bind 50% of the total protein including: Euphorbia hirta which has 3 compounds with 4 target proteins and has a percentage of the target protein is 80%, Mangifera indica which has 2 compounds with a target protein of 4 proteins and has a percentage of the target protein is 80%, Capsicum frutescens which has 3 compounds with a target protein of 3 proteins and has a percentage of the target protein is 60%, Carica papaya which has 3 compounds with 3 target proteins and has a percentage of the target protein is 60%, and Trigonella foenum graecum which has 1 compound with 3 target proteins and has a percentage of the target protein is 60%. In the combination of 2 plants, 14 formulas were obtained where all of these formulas targeted all the protein interactions obtained, so that the percentage value for all proteins was 100%..

INTRODUCTION

Hypertension can be defined as a disease characterized by a systolic blood pressure value of at least 140 mmHg and or a diastolic blood pressure value of at least 90 mmHg[1]. Broadly speaking, hypertension is divided into 2 types. In some patients with systemic arterial hypertension consistent with a specific known cause (eg, renal disease, endocrine disease, constriction of the heart aorta) it is also known as secondary hypertension. Most patients have essential hypertension or so-called primary hypertension which poses an increased risk of vascular disease / blood vessels (such as: thrombosis or rupture of blood vessels / stroke, heart infarction)[2], [3]. According to data obtained by Riskesdas, the higher the age, the higher the prevalence of hypertension in Indonesia. According to Riskesdas data, it is also known that the female gender is more susceptible to hypertension than men with a percentage value ratio of 36.9% for women and 31.3% for men [2]. The influence of the area of residence is known to contribute to the prevalence of hypertension, such as in urban areas the percentage value of hypertension is 34.4% and in rural areas is 33.7%. [4]

Hypertension is a condition when the blood pressure in the blood vessels is chronically elevated. This can happen because the heart works harder to pump blood to meet the body's needs for oxygen and nutrients. If left unchecked, this disease can interfere with the function of other organs, especially vital organs such as the heart and kidneys. The hypertension criteria used refer to the JNC VII 2003 diagnostic criteria, namely the results of measuring systolic blood pressure 140 mmHg or diastolic blood pressure 90 mmHg. (Kasper et al, 2015) The division of blood pressure levels according to has 7 types, namely: optimal (systolic blood pressure <120 mmHg and diastolic <80 mmHg), normal (systolic blood pressure 120-129 mmHg and diastolic 80-84 mmHg), high normal (systolic blood pressure 130-139 mmHg and diastolic 85-89 mmHg), high normal grade 1 (systolic blood pressure 140-159 mmHg and diastolic blood pressure 90-99 mmHg), high normal grade 2 (systolic blood pressure 160- 179 mmHg and diastolic 100-109 mmHg), high normal grade 3 (systolic blood pressure 180 mmHg and diastolic 110 mmHg) and isolated systolic hypertension (systolic blood pressure 140 mmHg and diastolic <90 mmHg)[5], [6]. Hypertension can be classified into 2 types, namely primary or essential hypertension (90% of cases of hypertension) whose cause is unknown and secondary hypertension (10%) caused by kidney disease, endocrine disease, heart disease and kidney disorders. According to the JNC VII Report 2003, the diagnosis of hypertension is made when a systolic blood pressure (TDS) 140 mmHg and/or a diastolic blood pressure (TDD) 90 mmHg is obtained on two measurements at different times [4], [7].

Tissue Pharmacology, based on the concept of multidisciplinary integrity, is a powerful tool for analyzing multi-level tissue from molecular-target-pathways-of-disease through interactions between traditional medicine and disease from a holistic perspective [8]. When searching for a drug with a computational pharmacy approach, the first step is to identify the function of the possible target and its role in the disease. The second step is target validation, in this step it is necessary to demonstrate that the target molecule is directly involved in the disease process. The next step is the identification of the active compound which is defined as the molecule that shows significant biological activity in the screening assay, i.e. the new material that links the chemical structure for target modulation. Until the last step, which is the 11th step which involves clinical testing of the drug concept that has been made. Based on this series of steps, target molecules (proteins) that play a role in the cause of a certain disease and active compounds of synthetic drugs will be obtained that function to target these molecules. The function of the active compound is to lock the target protein in order to cure a disease, or both are termed as “lock and key” [9], [10].The search on tissue pharmacology this time uses the Graph mining search method where the relationship between plants, active compounds and target proteins will be built which is presented in the form of graphs. This graph is formed from 2 types of data, the first data used to build the graph is training data which will be tested using test data. After this process, the data will be processed using machine learning which will create the final data for the graph. The graph that is formed will be analyzed based on several predetermined parameters[11]. This research was conducted with the following objectives 1. Looking for Indonesian plants that have an antihypertensive effect by using Tissue Pharmacology, 2. Looking for a plant combination formula that has high effectiveness as an antihypertensive by using Tissue Pharmacology, 3. Testing the potential of antihypertensive efficacious substances in plants by tissue pharmacology, 4.Building a pharmacological network related to compound-protein and protein-protein interactions, 5. Improving the science of Indonesian medicinal plants in Tissue Pharmacology. As for the benefits of this research for researchers, they can find out various kinds of Indonesian plants that have the potential as antihypertensives. For the community, it is hoped that this research can be useful as an alternative in developing the use of natural medicines as antihypertensive therapy. For advanced students, it is hoped that the testing in this study can be continued by in vitro and in vivo tests.

METHOD

The materials used in this study include: search for proteins related to hypertension at Uniprot, download results of interaction designs of compounds and proteins on STRING, cluster results from the Cytoscape application, download results of potential substance data on KNApSAcK, results of data downloads information on substances in PubChem, results of RStudio processing, results of data searches, extraction and labeling from Jupyter Lab, results of downloading of interaction designs of compounds and proteins on STITCH, which will later be analyzed and combined using the Orange application.

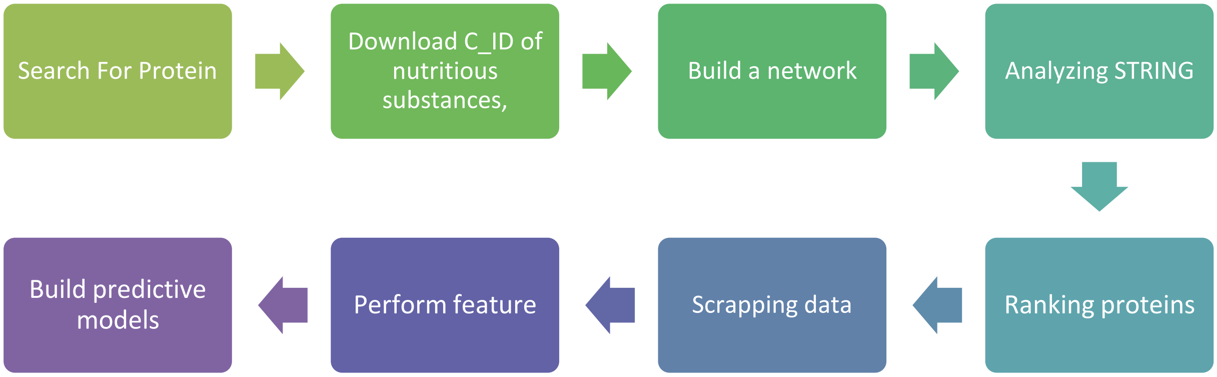

FIGURE 1 Working Method

Search for protein data related to hypertension

Proteins related to hypertension can be found and searched at https://www.uniprot.org/ for download. The list of proteins that appear on the web needs to be filtered because proteins related to hypertension have not only been tested on humans but also on animals so that animal proteins are also available on this web.

Potential Substance Download on KNApSAcK

The plants whose journals have been obtained are sought for additional information related to the effects of the nutritious substances listed in the database and the molecular formula code for the nutritious substances (C_ID). Downloading this potential substance, the website used is http://www.knapsackfamily.com, when on the web select the "metabolite activity" menu then click on the "Partial (Metabolite Name, Activity Category, Biological Activity and Target Species)" section. Enter the names of the nutritious substances that have been obtained from the data mining results in the "Metabolite Name" column one by one then check the box next to the column and press "enter". Information appears related to the efficacious effect of the substance, then click on the C_ID section of the substance to save the C_ID code and SMILES of the efficacious substance.

Protein and Protein Interaction Network Design (STRING)

The list of proteins that have been obtained from the screening stage was built to design a network of interactions between proteins that are thought to affect each other. Building a network design of this protein-protein interaction can be done by accessing the online web https://string-db.org. The data in it can be the results of in vitro or in vivo testing by researchers that have been entered into the web database. Select the multiple protein menu then enter the list of proteins that have been obtained on the SWISSTargetPredcition web, in the organism column select "Homo sapiens" then select search. Displayed results in the form of points (nodes) and lines (edges) with the names of the proteins in them and the interacting substances. Settings to determine the confidence value can be made on this website in order to minimize the potential for interactions that are considered small. After that the file is saved in the form of "TSV".

Compound and Protein Interaction Network Design (STITCH)

After obtaining a list of known compounds, we need to know what interactions occur between the active compounds and the proteins on which the compounds work. We can search for this information using the online web, namely http://stitch.embl.de/. The data in it can be the results of in vitro or in vivo testing by researchers that have been entered into the web database. This test step is to enter the STITCH web then select the multiple protein menu and enter a list of all the compounds that we got earlier. In the organism column select "Homo sapiens" then select search. Displayed results in the form of points (nodes) and lines (edges) with the names of the proteins in them and the interacting substances. Settings to determine the confidence value can be made on this website in order to minimize the potential for interactions that are considered small. After that the file is saved in the form of "TSV".

Clustering of protein interactions in Cytoscape

Cytoscape functions to integrate, analyze and design the form of interaction networks between proteins, compounds, and proteins. The Cytoscape application can be downloaded on the official web https://cytoscape.org with the latest version being 3.8.2. Import data in cytoscape, namely a file in the form of "TSV" which is obtained when analyzing using STRING. Select the Apps menu, install CytoCluster, make adjustments to the number of nodes and density in CytoCluster, after obtaining the results of several clusters, then to analyze the network, select the analyze menu available in Tools in the menubar. Columns that have been analyzed are deleted, except for the Betweeness Centrality, Degree, Stress and Closeness Centrality columns.

Calculation of Overall Centrality using RStudio

After obtaining data related to the clusters formed or the best from Cytoscape, the next step to do is to calculate the overall centrality of the clusters formed. To do this activity, RStudio is used because RStudio focuses on graphing and statistical processing so that we can enter commands to remove unnecessary parts and calculate the eigenvalues which we will get the overall centrality value. RStudio can be accessed by downloading anaconda navigator then installing RStudio in the application.

Extracting, Scrapping, and Labeling data using Jupyter Lab

The data from the overall centrality calculation need to be extracted both on the protein and the compounds that act on the protein. The protein that was obtained in the cluster was searched for substances that act on the protein through an automatic scraping method by entering a command in the Jupyter Lab. The protein obtained will be extracted in the form of a dipeptide while the compound obtained will be extracted in the form of a fingerprint. After all the compounds and proteins are combined in the Jupyter Lab, data labeling is carried out so that data that has interactions can be distinguished.

Merging data that has been processed and analyzed using machine learning Orange

After going through the extraction, scraping, and labeling processes, the data generated from these processes need to be combined and analyzed. Data that has been labeled will be processed and analyzed to read the label that has been given in the previous process[12]. The result of this analysis is the label that was previously worthless becomes valuable and provides an indicator regarding the presence or absence of interactions between proteins and compounds. After that, machine learning is tuning for modeling in order to find out which data model is the best and the data specificity, sensitivity and precision of the data can be known

RESULT AND DISCUSSION

The search process for significant proteins related to hypertension is carried out in the Uniprot database on the initial web display there is a search column, the keyword "Hypertension" is entered in the column. There is a display of protein search results related to hypertension select on the left that reads "Homo sapiens" because the search for hypertension protein is only limited to proteins found in humans. The results of this search obtained as many as 270 proteins in the human body related to hypertension, then this data was downloaded and stored.

The list of proteins that was obtained from the Uniprot database was then searched for interactions between proteins using the STRING web (Search Tool for the Retrieval of Interacting Genes/Proteins). The initial display on the web is selected in the "Search" section then selects "Multiple Proteins" and entered the list of proteins that were obtained then set the organism to "Homo sapiens" and selected "advanced settings" set "required score" to the highest confidence. The next screen will inform whether or not the protein is being searched for and its interaction on the STRING web is selected to continue, then interactions of all proteins will appear and this interaction will then be downloaded and saved in the form of a TSV (Tab Separated Values) file so that this file can be read in the Cytoscape application. . The results of the search for these protein-protein interactions contained a total of 248 interactions.

The stages of pre-processing the data begin with entering the results of the protein interaction into the Cytoscape application by selecting the file menu then selecting import then selecting network from file then selecting the protein interaction file in the form of "TSV" then selecting ok. The next screen will display a setting column to form a network by entering certain assessment elements, select select none then the first column of the protein interaction file is changed to "source node" and the second column is changed to "target node" then select ok. The network formed is known to have a total of 136 nodes and a total of 248 edges. The next process is to cluster the interactions of these proteins in order to find out which protein groups are significant or have high interactions. This clustering stage uses the CytoCluster feature in Cytoscape by selecting the apps menu then selecting CytoCluster then selecting the ClusterOne section. There are 2 settings made on ClusterOne, the first is the Basic Parameter where the minimum size is set to 10 then the minimum density is set to 0.5 then the second setting is the advanced parameter where the seeding method is set to from every node then select ok. The results of this process are obtained as many as 4 clusters which will later be analyzed one by one. How to take each cluster using the file feature then new network then select the selected nodes form, selected edges then select the Tools menu then analyze network. The results of this network analyze then that are taken are 4 parameters, namely: Betweeness centrality, Closeness centrality, Degree, and Stress, later these four parameters will be processed in Python to calculate overall centrality in order to determine the significant protein.

The calculation of overall centrality is done by entering a script/command in the kernel in the Python program, The results obtained are that the AGTR1_HUMAN protein has an overall centrality value of (2.76967082423431), the GNAS_HUMAN protein has an overall centrality value of (-1,38194807757488), ACE2_HUMAN protein has an overall centrality value of (-1,72856432286345), TAC3_HUMAN protein has an overall centrality value (-1.01621749330442), EDN1_HUMAN protein has an overall centrality value of (-0.392822941980431). These values indicate how significant a protein is in a cluster.

TABLE 1 Overal Centrality Protein

No	*Protein*	*Overal Centrality*
1	AGTR1_HUMAN	2,76967082423431
2	GNAS_HUMAN	-1,38194807757488
3	ACE2_HUMAN	-1,72856432286345
4	TAC3_HUMAN	-1,01621749330442
5	EDN1_HUMAN	-0,392822941980431

Data analysis is done by tracing the graph (graph mining) of each data created. The first data is searches related to compounds that interact with proteins, obtained from a total of 5740 data as many as 535 data provide "interaction" results between compounds and proteins, while as many as 5205 data provide "no interaction" results between compounds and proteins as shown in Table 1. Obtained from 535 interaction data as many as 5 proteins that are the target of this protein compound, namely ACE2_HUMAN, AGTR1_HUMAN, EDN1_HUMAN, GNAS_HUMAN, and TAC3_HUMAN, it is known that these proteins lead to proteins that cause hypertension. Next is a search related to compounds and plants, it is known that compounds that interact with protein are obtained as many as 119 compounds and after being traced, it is found that 76 plants have related compounds as seen in Table 1. Next, a network is built. The purpose of the network is to see a combination of interactions, both interactions between proteins with compounds and compounds with plants so that they form 1 network. Then an assessment is carried out on 1 plant related to the percentage of protein targeted, this calculation is carried out using the Python program. The assessment of this analysis is to look for plants that bind to 50% of the total protein that is the target protein. The results of this analysis can be seen in Table 1. Next, the search for the best combination of plants is carried out, a combination of plants is used a combination of 2 plants. The search for this combination of plants uses the Python program using graph mining method.

The results of the search for the best 1 plant found 5 plants that can bind 50% of the total protein including: Euphorbia hirta which has 3 compounds with 4 target proteins and has a percentage of the target protein is 80%, Mangifera indica which has 2 compounds with a target protein of 4 proteins and has a percentage of the target protein is 80%, Capsicum frutescens which has 3 compounds with a target protein of 3 proteins and has a percentage of the target protein is 60%, Carica papaya which has 3 compounds with 3 target proteins and has a percentage of the target protein is 60%, and Trigonella foenum graecum which has 1 compound with 3 target proteins and has a percentage of the target protein is 60%. In the combination of 2 plants, 14 formulas were obtained where all of these formulas targeted all the protein interactions obtained, so that the percentage value for all proteins was 100%.

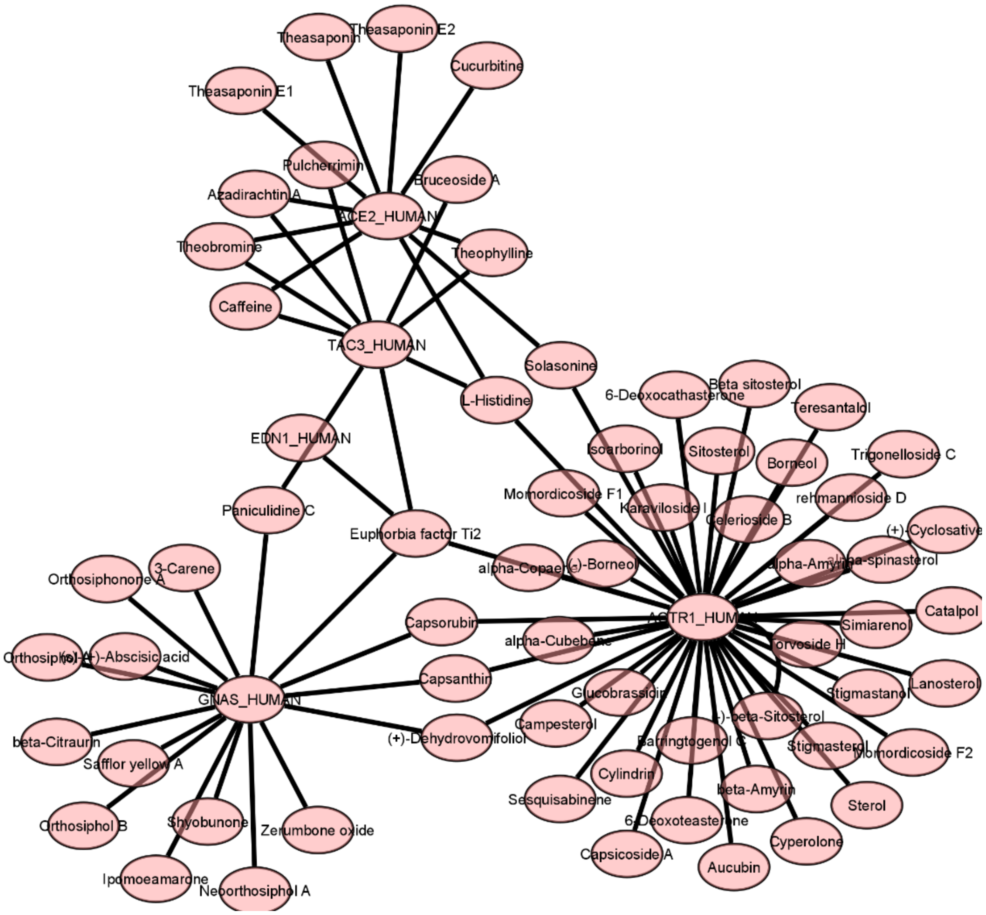

FIGURE 2 Protein Prediktion

CONCLUSION

The graph mining technique in the prediction of herbal formulas this time produced 14 formulas with a target protein percentage of 100%. The formula obtained is produced from 2 combinations of plants that are predicted to have compounds that interact with proteins that cause hypertension. Explanations related to the pathway of proteins that cause hypertension can be found by tracing using CytoKegg. In the results of 1 plant which is considered the best plant, 6 plants have a percentage score of protein above 50%. The suggestions from the research on making tissue pharmacology that have been carried out are, add protein data from other databases to increase target proteins in tissue pharmacology. Added plant combinations to look for formula diversity, Use of different methods in determining the best formula such as calculating edge weights on nodes, crop reduction speed and so on, and In vivo testing to provide information regarding the validity and effectiveness of the predicted formula.

REFERENCES

1 G. Yulanda and R. Lisiswanti. Majority **6(1)** 25–33 (2017)
2 B. Nuraini. J Major **4(5)** (2015)
3 F. H. Messerli, B. Williams, and E. Ritz. Lancet **370(9587)** (2007)
4 K. Makó, C. Ureche, and Z. Jeremiás. J. Cardiovasc. Emergencies **4(2)** (2018)
5 S. Mahmood *et al.* J. Med. Sci. **188(2)** 437–452 (2019)
6 F. Huwaina *et al.* Journal Fildza Huwaina Fathnin **5(3)** (2020).

7 Y. L. Cuffee *et al.* Journal Hypertens **32(2)** (2019)
8 Y. C. Hung, S. M. Huang, Q. P. Lin, and M. L. Tsai. Data Syst.(2005)
9 M. Wozniak, B. Michalak, J. Wyszomierska, M. K. Dudek, and A. K. Kiss. Front. Pharmacol. **9(APR)** (2018)
10 A. Parveen, S. Choi, J. H. Kang, S. H. Oh, and S. Y. Kim. Int. J. Mol. Sci. **22(1)** (2021)
11 T. Slimani *et al.* Comput. Electron. Agric. **2(1)** 61–67 (2015)
12 J. Violos, K. Tserpes, I. Varlamis, and T. Varvarigou. Front. Appl. Math. Stat. **4** (2018)