Indonesian Journal of Biotechnology

VOLUME 26(4), 2021, 197-205 | RESEARCH ARTICLE

Analysis using top-k skyline query of protein-protein interaction reveals alpha-synuclein as the most important protein in Parkinson's disease

Mohammad Romano Diansyah[1], Annisa[2], Wisnu Ananta Kusuma[1,2*]

[1]Department of Computer Science, IPB University, Jln. Meranti, Kampus IPB Darmaga, Bogor 16680, Indonesia
[2]Trophical Biopharmaca Research Center, IPB University, Jl. Raya Dramaga, Kampus IPB Dramaga, Bogor 16680, Indonesia
*Corresponding author: ananta@apps.ipb.ac.id

SUBMITTED 10 January 2021 **REVISED** 4 September 2021 **ACCEPTED** 18 October 2021

ABSTRACT Parkinson's disease is the second-most-common neurodegenerative disorder and can reduce patients' quality of life. The disease is caused by abnormalities in dopaminergic neurons, such as reactive oxygen species (ROS) imbalance leading to programmed cell death, protein misfolding, and vesicle trafficking. Protein-protein interaction (PPI) analysis has been demonstrated to understand better candidate proteins that might contribute to multifactorial neurodegenerative diseases, particularly in Parkinson's disease. PPI analysis can be obtained from experiments and computational predictions. However, experiment data is often limited in interactome coverage. Therefore, additional computational prediction methods are required to provide more comprehensive PPI information. PPI can be represented as protein-protein networks and analyzed based on centrality measures. The previous study has shown that top-k skyline query, a method using dominance rule-based centrality measures, reveals important protein candidates in Parkinson's diseases. This study applied the top-k skyline query to PPIs containing experiment and prediction data to find important proteins in Parkinson's disease. The result shows that alpha-synuclein (SNCA) is the most important protein and is expected to be a potential biomarker candidate for Parkinson's disease.

KEYWORDS centrality measures; Parkinson's disease; significant protein; top-k skyline query

1. Introduction

Parkinson's disease (PD) is a disease that can be recognized by several symptoms which may appear, such as decreased motor functions, autonomic dysfunction, hallucinations, and depression (DeMaagd and Philip 2015). As the disease may worsen and cause pneumonia, it can threaten the patients' life. Furthermore, the disease can lower patients' quality of life and impact their families and society (DeMaagd and Philip 2015). The disease burden was estimated to rise from 4.1 to 4.6 million in 2005 to 8.3 to 9.3 million in 2030 (Dorsey et al. 2007), which may broadly impact crowded nations, particularly several Asian countries such as China, India, and Indonesia. Currently, PD has been known as one of the most common neurodegenerative disorders with incidence ranging from 16 to 19 per 100,000 people per year (Twelves et al. 2003; Lebouvier et al. 2009; WHO 2004) and expected to overcome cancer as the second most common cause of death in 2040. Furthermore, the economic burden of PD direct and indirect cost of treatment reached US$ 1,100 million worldwide (Twelves et al. 2003; WHO 2004).

PD is a disease known as neuron dysfunction. It mainly impacts dopaminergic receptors due to several factors such as reactive oxygen species (ROS)-induced cell death (Dias et al. 2013), protein misfolding (Tan et al. 2009), or changes of proteins that are responsible for vesicle trafficking (Esposito et al. 2012), G protein activations (Odagaki and Toyoshima 2006), and many proteins which should be noticed carefully. Proteins interact with each other in carrying out their function and often called protein-protein interaction (Chang et al. 2016). Protein-protein interaction (PPI) is a good representation for unraveling protein functions, disease-disease, and disease-gene associations (Liu et al. 2015; Chang et al. 2016). Therefore, the PPI analysis to predict significant protein candidates that play a role during the disease progression provides a better understanding of multifactorial degenerative diseases, including PD.

Currently, many databases store PPI information, such as STRING. STRING database (string-db.org) is a PPI database with the largest number of organisms and proteins (Szklarczyk et al. 2018). The database provides two types of interaction. The first one is experimental data obtained from experiments. The second type is prediction interaction data obtained from many methods, including co-expression analysis, detection of shared selective sig-

nal across genomes, text-mining, and computational transfer knowledge based on gene ontology (Szklarczyk et al. 2018). STRING's experimental proteins interaction information was collected from other databases such as BIND, DIP, GRID, HPRD, IntAct, MINT, and PID.

PPI analysis is often limited by interactome coverage, where interactome is a set of PPI that can occur inside a cell (Yu and Fotouhi 2006). The interactome coverage is a ratio between PPI that occurred inside the cell and interactome often stated in percentage (%). For example, human is predicted to have 650,000 PPI (Stumpf et al. 2005). However, Human Protein Reference Database (HPRD) (https://hprd.org) , accessed in December 2019, only has 41,327 PPI information covering 6.3% interactome. Experimental data can be combined with prediction data To improve interactome coverage (Jansen et al. 2002; Liu et al. 2015).

The PPI network can be represented as a graph with proteins as nodes and interactions as edges. The measure of centrality can be applied for finding the subnetwork, even the importance of a node in a network. Thus, data transformation can be done from a graph to an object with centrality measures as attributes. However, there were many centrality measures with different characteristics, which led to debate among the researchers to determine which centrality measures are better (Raman et al. 2014).

In PPI analysis, clustering is frequently used to predict proteins function (Hao et al. 2016). Previously, several studies focused on centrality measures and machine learning were conducted to reveal PPIs subnetworks that have an important role in certain diseases such as Diabetes (Usman et al. 2019). In this study, we try to better understand which proteins play a significant role in PD. Previously, Diansyah et al. performed the Skyline Query to predict PPI in PD (Diansyah et al. 2019). In this study, we performed Skyline Query, an algorithm for finding non-dominated data, along with centrality measure to find significant proteins of PD. Skyline query (SQ) is an algorithm that shows the optimal solution for the problem with various criteria based on dominance rules (Borzsonyi et al. 2001). This algorithm is developed based on the maximal vector problem in mathematics. The result of SQ is a set of non-dominated objects called Skyline Objects. An object dominates another object only if it has the same score or a better score in all attributes and better at least in one attribute (Borzsonyi et al. 2001). Commonly, SQ is used to find the optimal object, for instance, a hotel or restaurant, that meets multiple conflicting criteria.

In this study, we employed SQ to find the significant proteins that have essential roles in the regulation of PD. The logic of finding skyline object is in line with finding significant proteins, which attribute values are not less than that of any other protein and has at least one attribute whose value is greater than that of any other protein. We employed top-k SQ, one of the variants of SQ, to overcome the weakness of SQ which is not robust against an increasing number of attributes. We used seven centrality measures, namely degree, betweenness, closeness, eigenvector, eccentricity, radiality, and bridging as attributes. We combined experiment data and prediction data to improve interactome coverage (Jansen et al. 2002).

2. Materials and Methods

We conducted this research in four stages. First, we collected the necessary data for this research. Second, we performed data pre-processing. This step included removing duplicate data, deleting unconnected networks, and transforming the network into centrality measures. Third, we applied the Top-k Skyline Query to find the significant proteins. Finally, we analyzed the results by conducting a literature review to determine whether the Top-k Skyline Query could be used to find the significant proteins. Figure 1 shows the flow chart of this research.

2.1. Dataset

We collected datasets from OMIM (https://omim.org/) and STRING database (https://string-db.org/) on March 11th, 2020. The OMIM database was used to find proteins associated with PD. Moreover, the STRING database was used to find the protein interaction associated with PD. The first step was to find protein associated with PD from OMIM. The query at OMIM was conducted using "+" as a prefix for every word. The prefix was used to get the precise

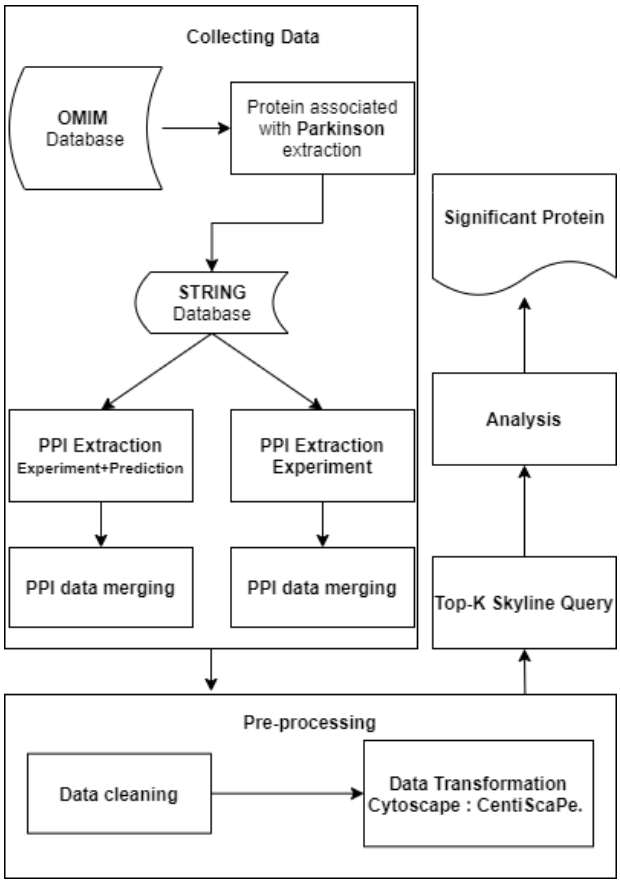

FIGURE 1 Flow chart of the study.

result. The query for this study in OMIM is "+Parkinson +Disease".

The second step is to find the PPI in STRING for proteins that we get from OMIM. For each protein associated with PD, there was a separate interaction file, so that we needed to combine the data into one file. This study was done by developing a program or scraper in Python 3.7 to automate this step. Figure 2 shows the pseudocode of data scraping. Moreover, in this study, we used the combination of the experimental dataset and prediction dataset from STRING.

```
Request Stringdb API (protein, string method)
Get interacction of proteins
    Sort node according to alphabet
    Remove redudant node  //remove redudant
protein
Remove duplicate interaction
    Sort interaction
    Remove duplicate
Get PPI
```

FIGURE 2 Pseudocode of data scraping.

2.2. Data pre-process

We used Cytoscape (https://cytoscape.org/) for conducting pre-processing data. Two main steps in this study include data cleaning and data transformation. First, we visualized the PPI data to find any unconnected network. A Network that was not connected to the main (biggest) network would be removed. We assumed that the significant proteins are located in the back bound network, a collection of nodes with a high number of members and a high density. Thus, the unconnected networks to the back bound were removed. Next, we omitted the duplicate interaction data. The last step was to transform the data from the protein network into centrality measures. This process was done by using CentiScaPe 2.2 in the Cytoscape application (Scardoni et al. 2009; Scardoni and Lau 2012). After data transformation was completed, proteins with seven centrality measures were exported into a comma-separated value file (csv). Next, the output was processed in further steps.

2.3. Centrality Measures

Centrality measures are a unit of measure to measure the important node in a network interaction and have been widely used for analysis in biological networks. Many centrality measures can be used to measure the importance of a node. In this study, seven values of centrality measures were used, namely degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, radiality, eccentricity, and bridging centrality.

Degree centrality is the simple calculation of centrality. Degree centrality is obtained by counting the number of edges connected to the node. The probability of a protein becoming the center of regulation is directly proportional to the greater degree of centrality (Scardoni and Lau 2012).

Betweenness centrality can be obtained by calculating the shortest path by adding the shortest path through the node divided by the total number of shortest paths in the graph. The greater the betweenness centrality value, the more likely the node is often bypassed for communication between proteins so that the more relevant it is to become a regulatory protein (Scardoni and Lau 2012).

The calculation for closeness centrality is based on the number of shortest paths from one node to another node. The value of the number of shortest paths is used as a divisor of 1. Thus, the greater the value of closeness centrality, the more central the position of the protein is. Therefore, it can become a regulatory protein for other proteins in the network (Scardoni and Lau 2012) Eigenvector centrality is calculated based on the concept that if a node-*i* is connected to another node with a high score, node- *i* will also have a high score (Scardoni et al. 2009). The initial step in finding eigenvector centrality is to find the largest eigenvalue first, then using the largest eigenvalue, the eigenvector matrix will be obtained. The eigenvector centrality value was obtained by dividing the eigenvector matrix of a node by the determinant value of the eigenvector matrix. The greater the eigenvector centrality value indicates if the node interacts with other important proteins to become a regulatory center for other important proteins (Scardoni et al. 2009).

Radiality is based on the shortest path from one node to another node. Before adding up, the shortest path value is used to reduce (Δ_G + 1) where Δ_G is the largest shortest path, after which it is added. The higher the radiality value of a node is functionally relevant to other nodes. The high values of radiality, eccentricity, and closeness centrality indicate the consistency of a node to become the center of the network (Scardoni and Lau 2012).

Eccentricity is calculated by finding the largest, the shortest path from one node to another node. According to Scardoni and Lau (2012), in biological terminology, eccentricity can indicate a protein's convenience reached by other proteins in the network. The greater the eccentricity value suggests that it is easy to influence other proteins in the network.

Bridging centrality is the result of the development of betweenness centrality. The bridging centrality value is obtained from the multiplication of the betweenness centrality and the bridging coefficient. A node with a high value of bridging centrality indicates if the node connects a node with a high degree to connect between clusters in the interaction network (Scardoni et al. 2009).

All values of centrality measures that have been described will be used as attributes for each protein. Furthermore, this data was used for the following process to select interesting objects based on seven criteria of centrality measures by using Skyline Query.

2.4. Skyline Query

Skyline query (SQ) is a method to find the non-dominated object; this algorithm chooses an interesting object from a dataset. An object is later categorized as an interesting object if not dominated by another object (Borzsonyi et al. 2001). For example, object A dominates object B if A has the same score or a better score in all attributes than B and better at least in one attribute (Borzsonyi et al. 2001). Then this rule in SQ is called the dominance rule.

In this study, the higher score in centrality measures means a higher chance of the protein being an important protein for every centrality measure. So, the dominance rule for Table 1 is the highest score in degree centrality and closeness centrality. The results of implementation SQ in Table 1 were the object A and C. Object B has the same score as object A in terms of degree centrality. However, it has a lower score in closeness centrality that makes object A dominates object B. Object D is dominated by A because it has the lowest score in every attribute compared to object A. Since no other object can dominate object A and C, object A and C became the skyline object for Table 1.

TABLE 1 Dataset example with two centrality measures.

Object	Degree	Closeness
A	30	0.0045
B	30	0.0015
C	50	0.0015
D	20	0.0035

However, SQ has a weakness: the more attributes that are used, the more skyline objects will be used so that the results are no longer relevant (Kontaki et al. 2008). This study used a developed SQ called top-k skyline query (top-k SQ). Top-k SQ ranks skyline results to find the most important data in skyline objects. The ranking is done by searching the most dominant data. This method finds data that dominates other data, and the most dominant data was in the top result. This study used centrality measures as attributes and top-k SQ to analyze PPI.

Based on the concept of top-k SQ, a protein is a protein that is not dominated by another protein with the order by how many proteins it dominated. The most important result of top-k SQ is a candidate for important proteins related to the disease that was later further cross-checked. Since there are many centrality measures, this study only used basic centrality measures and the other two centrality measures. The basic centrality measures in graph theory are degree, betweenness, closeness, eigenvector, and eccentricity (Sharma et al. 2016). In this study, besides the basic centrality measures, we used radiality and bridging centrality.

We used top-k SQ to find an important protein of PD using seven centrality measures (degree, betweenness, closeness, eigenvector, eccentricity, radiality, and bridging). There are two interactions data types based on their resources, experiment data and experiment+prediction data. We used experimental data to determine whether interactome coverage in PD good enough for PPI analysis. This study used SQ, an algorithm for finding non-dominated data, and centrality measure to find important proteins of PD.

2.5. Top-k Skyline Query

Top-k representative skyline query (top-k RSP) is a top-k SQ algorithm used to maximize data dominated by k skyline objects (Lin et al. 2007). The complexity for top-k RSP is $O(kn^2+kn)$, where n is the total number of data. This study chose a basic top-k SQ because the data is relatively small, and the process is done only once. Figure 3 shows the pseudocode of top-k RSP.

```
N = Input Data                                          #read
  input
top_k  = {}                                             #output data
from 1 to k
   find skyline with highest domination score from N
   sky = skyline with highest domination score
   find data dominated by sky
   dom = data dominated by sky
    top_k = append(top_k,sky)
    remove sky and dom from N
end
```

FIGURE 3 Top-k RSP pseudocode.

Using SQ, the skyline objects of data in Table 1 is object A and object C. The object of D is dominated by object A which has a better score in all dimensions than object D. Object B is dominated by object A because it has a lower score in closeness centrality. However, it has the same degree centrality score. Object C dominates object B because it has the same score in closeness with a better score in degree. Object A and object C is incomparable because object A has a better score of closeness centrality; otherwise, object C has a better degree of centrality. No other data could dominate objects A and C, so objects A and C are skyline objects.

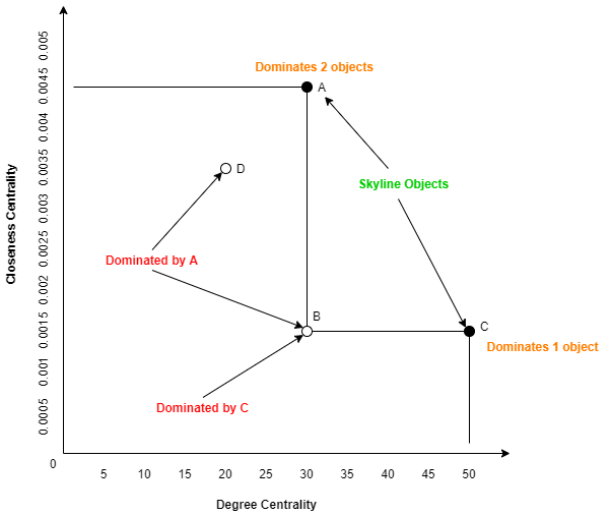

FIGURE 4 Visualization for top-k skyline query

Top-k SQ ranks the skyline objects by how much data

were dominated by the skyline objects. As shown in Figure 4, object A dominates two data (D and B) while object C only dominates one data (B). Object A is the highest rank for top-k SQ because it dominates the most. Therefore, the top-k SQ for Table 1 is A, and the top-k SQ for Table 1 is A and C.

2.6. Data Analysis

The objective of this step is to analyze the result of top-k SQ. The proteins relations to PD were cross-checked with the experimental data, particularly the highest rank skyline object we get from top-k SQ. Further analysis would define whether experiment and experiment + prediction data can be used in the PD PPI analysis. We expect to see the effect of interactome coverage.

3. Results and Discussion

There were 271 proteins data related to PD obtained from OMIM, but only 252 proteins have interaction information in STRING. Therefore, proteins associated with PD which are not found in STRING were excluded from this study. Two hundred and fifty-two protein interaction files were merged into one for each interaction source. Table 2 shows the results from STRING after merging the interaction files.

From Table 2, there are 1,553 proteins with 4,868 interaction data with interaction source only from the experiment. Meanwhile, there are 1,848 proteins with 8,577 interactions from experiment and prediction interaction sources. Visualization of experimental data can be seen in Figure 5. Figure 5 shows many unconnected networks. Networks that are not connected to the main graph were deleted. After the deletion of the unconnected graph, data duplicate will be removed as well. Figure 5 shows the visualization for experiment data after data cleaning. In Figure 5 and Figure 6, the red nodes represent proteins associated with PD that we obtained from OMIM. Meanwhile, the green nodes represent protein without direct association with PD (interaction protein from STRING).

TABLE 2 Results from STRING after merged.

Interaction Source(s)	Number of Proteins	Number of Interaction
Experiment	1,553	4,868
Experiment + Prediction	1,848	8,577

However, deletion in duplicate data and unconnected networks will decrease the number of proteins and interactions. Table 3 shows the number of proteins and interactions before the protein networks were transformed into centrality measures. From Table 3, there are only 1,269 proteins and 4,198 interactions for the experiment data interaction source. Moreover, 1,682 proteins with 7,894 interactions left for the experiment+prediction data source.

TABLE 3 Results from STRING after data cleaning.

Interaction Source(s)	Number of Proteins	Number of Interaction
Experiment	1,269	4,198
Experiment+Prediction	1,682	7,894

After data cleaning, PPI networks were transformed into centrality measures using CentiScaPe 2.2. There are two default outputs: a name and a shared name. Since

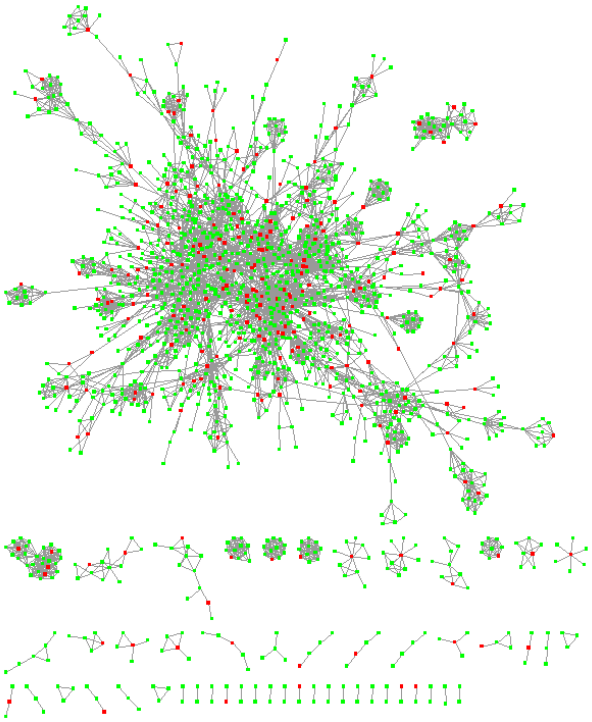

FIGURE 5 Experiment data visualization before data cleaning.

FIGURE 6 Experiment data visualization after data cleaning.

both contain the same protein name, the shared name was omitted. The transformation data results are proteins with seven centrality measures as attributes with one protein name (name, degree centrality, betweenness centrality, closeness centrality, eccentricity, eigenvector centrality, bridging centrality, and radiality). The data is transformed, then exported into a comma-separated value (CSV) as the input for top-k SQ.

Data with interaction source experiment is processed first. The maximum k for top-k SQ is 21 since there are only 23 skyline objects resulting from SQ. Proteins included in the top-21 SQ were SNCA (alpha-synuclein), PARK2 (parkin RBR E3 ubiquitin-protein ligase), TRAF2 (TNF Receptor Associated Factor 2), FN1 (Fibronectin 1), HSPA8 (Heat Shock Protein Family A (Hsp70) Member 8), GPR37 (G-Protein Coupled Receptor 37), TRAPPC1 (Trafficking Protein Particle Complex Subunit 1), LRRFIP1 (LRR binding FLII interacting protein 1), TH (Tyrosine Hydroxylase), GLB1 (Galactosidase beta 1), CTSA (Cathepsin A), PTPRC (Protein Tyrosine Phosphatase Receptor Type C), GSR (Glutathione-Disulfide Reductase), TNF (Tumor Necrosis Factor), C4BPA (Complement Component 4 Binding Protein Alpha), PRNP (prion protein), TP53 (Tumor Protein 53), MAPK8 (Mitogen-Activated Protein Kinase 8), SOD1 (superoxide dismutase 1), HSP90AA1 (Heat Shock Protein 90 Alpha Family Class A Member 1), and GNAI2 (G Protein Subunit Alpha I2). Table 4 shows the biological and experimental associations of genes with PD.

However, the top-1 SQ is SNCA since SNCA dominates most data. The result for the top-3 SQ for experimental data can be seen in Table 5. SNCA is the most important protein because it dominates another protein (1,217 proteins). Meanwhile, the other protein dominates only vary from 0-14 proteins.

The following process used data interaction sources were experiment and prediction. Maximum k for top-k SQ for experiment+prediction data is ten since there were only ten skyline objects. The results were SNCA, TP53, KNG1 (Kininogen-1), PRDX5 (Peroxiredoxin-5), GTPBP4 (GTP Binding Protein 4), PABPC1 (Polyadenylate-binding protein 1), ANXA1 (Annexin A1), AKT1 (RAC-alpha serine/threonine-protein kinase), PARK2, and APP (Amyloid Beta Precursor Protein). Aside from SNCA, TP53, and PARK2, relations to PD was shown in Table 6. In this table, we added further information on other genes that were not mentioned in the previous table.

Among ten skyline objects, the most important protein is SNCA. SNCA results from a top-1 SQ; it means that SNCA dominates another protein. Table 3 shows the result for top-3 SQ with experiment+prediction data as the interaction source. Based on Table 5, SNCA dominates 1663 another protein, so that it becomes the most important skyline object based on top-k SQ.

The execution time for the Python program is 0.2532 s for experimental data and 0.1508 s for experiment+prediction interaction data. Since both data types

TABLE 4 List of proteins that related to Parkinson disease's.

Proteins	Association to Parkinson's disease (PD)
GPR37	Highly expressed in neuronal progenitor cells, in particular Wnt-dependent neurogenesis (Berger et al. 2017)
GNAI2	Expression is increased during stress and plays important role to inhibits adenylate cyclase, to modulate cAMP mediated responsed beta adrenergic stimuli (Tsolakidou et al. 2010)
SNCA (alpha-synuclein)	Located in presynaptic terminals and critical to regulate neurotramsiter release and vesicle trafficking (Mata et al. 2010)
	Commonly detected in Lewy bodies, which known as pathologic features of PD (Siddiqui et al. 2016)
PARK2	Controls program cell death and apotptosis (Konovalova et al. 2015)
	PARK2 germline mutations leading to cause neurons dysnfunctions (Veeriah et al. 2010)
	Mutations caused imbalance of program cell death and increase apoptosis (Konovalova et al. 2015)
TH (Tyrosine Hydroxilase)	An enzyme in dopamine biosynthesis. TH expression is foundly related to occurence of PD (Chen et al. 2017)
HSPA8	Decreases during aging and may postulated to PD's, which may affect autophagy process due to response of ER stress by protein unfolding (Loeffler et al. 2016)
TRAF2	Overexpression of TRAF2/6 may induced by chronic inflammations and hypothized to be reason of occurence PD (Chung et al. 2013)
TP53	One of the disease hallmark (Szybinska and Lesniak 2017)
SOD1	SOD1 proteinopathy known as neurotoxic superoxide dismutase 1 (SOD1). SOD1-associated familial amyotrophic lateral
	sclerosis (fALS) is recapitulated in idiopathic PD (Trist et al. 2018)
TRAPPC1	Not found
GLB1	Not found
HSP90AA1	Not found
FN1	Not found
MAPK8	Downregulated and a possible biomarker of PD (Chi et al. 2018)
C4BPA	Not found
PTPRC	Expression in blood is downregulated in PD (Bottero et al. 2018)
CTSA	Not found
PRNP	Not found
LRRFIP1	Not found

TABLE 5 Result of Top-3 Skyline Query for experiment data.

Protein	Dominates
SNCA	1,217
PARK2	14
TRAF2	11

TABLE 6 List of additional proteins that related to Parkinson's disease.

Proteins	Association to Parkinson's disease (PD)
KNG1	Level in cerebbrospinal is a potential marker of cognitive impairment in PD (Markaki et al. 2020)
PRDX5	Not found
GTPBP4	Not found
PABPC1	Not found
ANXA1	Not found
AKT1	Involved in protection against PD (Xiromerisiou et al. 2008)
APP	Not found

return the same protein that is SNCA as the important protein, there is only one candidate for important protein. Among those proteins, at least five proteins were found related to the PD. One of the most important proteins was alpha-synuclein (SNCA). Biologically, SNCA was responsible for presynaptic terminals and critical to regulating neurotransmitter release and vesicle trafficking (Mata et al. 2010). In addition, Alpha-synuclein is commonly detected in Lewy bodies, which is known as pathologic features of PD (Siddiqui et al. 2016).

TABLE 7 Result of top-3 skyline query for experiment+prediction data.

Protein	Dominates
SNCA	1,663
TP53	6
KNG1	3

Besides SNCA, several proteins are important in disease progressions; for instance, *PARK2* controls programmed cell death and apoptosis (Chen et al. 2017). PARK2 germline mutations are the leading cause of neuron dysfunctions (Chi et al. 2018). *PARK2* mutations caused an imbalance of programmed cell death and increased apoptosis (Konovalova et al. 2015). In GPCR classes, the GRP37 gene is highly expressed in neuronal progenitor cells, particularly Wnt-dependent neurogenesis (Berger et al. 2017). GNAI expression increases during stress and plays an important role in inhibiting adenylate cyclase, modulating cAMP, and mediating responses to beta-adrenergic stimuli (Tsolakidou et al. 2010). Lastly, tyrosine hydroxylase (TH) is an enzyme in dopamine biosynthesis, and since PD is related to dopaminergic neurons, TH expression is also related to the occurrence of PD (Chen et al. 2017).

SNCA is the first gene linked to PD. SNCA itself is thought to have an essential role in synaptic transmission (Mata et al. 2010). This protein has been given an identification name to show that SNCA is linked to PD and plays an important role: PARK1 (Klein and Westenberger 2012). SNCA is considered involved in the early onset of familial Parkinson's disease (FPD) as a major causative gene. It has been identified five mutations point in SNCA that cause autosomal dominant Parkinson's (Siddiqui et al. 2016). A study by Diansyah et al. (2019) found 14 proteins resulting from a skyline query in PD, and SNCA is one of the results. However, it still lacks information about the most important protein to PD. This study shows that SNCA is the most important protein for PD.

Experiment and experiment+prediction data give the same result proving its significance. It shows that this method can use experimental data in the PPI analysis for PD. Also, it indicates that interactome coverage in PD is good enough for PPI analysis since experiment data give an important protein as the result of this method. However, we need to do extended research to prove that interactome coverage in PD is sufficient.

4. Conclusions

Based on the result of this study, it can be concluded that the top-k skyline query can be used to find important proteins in Parkinson's disease (PD). Experiment and experiment+prediction interaction data sources for PD can be used in PPI Analysis using this method. The important protein for PD based on this study is alpha-synuclein (SNCA) that has been proven to have a significant role in this disease.

Acknowledgments

This research is supported by the Ministry of Research, Technology and Higher Education, Indonesia, under Master's Thesis Research Grant from Directorate of Higher Education, Indonesia, 2020. In addition, we thank Aryo Tedjo from the Department of Medical Chemistry, Faculty of Medicine Universitas Indonesia, and Dimas Ramadhian from Human Cancer Research Center, Indonesia Medical Education and Research Institute, Faculty of Medicine Universitas Indonesia for adding proteins' information.

Authors' contributions

MRD, ANN, WAK designed the study. MRD carried out the implementation work. MRD, ANN, WAK analyzed the results. MR wrote the manuscript. ANN, WAK revised the article and approved the final version of the manuscript.

Competing interests

The authors declare there is no conflict of interest.

References

Berger BS, Acebron SP, Herbst J, Koch S, Niehrs C. 2017. Parkinson's disease associated receptor GPR 37 is an ER chaperone for LRP 6 . EMBO Rep. 18(5):712–725. doi:10.15252/embr.201643585.

Borzsonyi S, Kossmann D, Stocker K. 2001. The Skyline Operator. In: Proceedings 17th International Conference on Data Engineering. p. 1–20. doi:10.1109/ICDE.2001.914855.

Bottero V, Santiago JA, Potashkin JA. 2018. PTPRC expression in blood is downregulated in Parkinson's and progressive supranuclear palsy disorders. J Parkinsons Dis. 8(4):529–537. doi:10.3233/JPD-181391.

Chang JW, Zhou YQ, Ul Qamar MT, Chen LL, Ding YD. 2016. Prediction of protein–protein interactions by evidence combining methods. Int J Mol Sci. 17(11). doi:10.3390/ijms17111946.

Chen Y, Lian Y, Ma Y, Wu C, Zheng Y, Xie N. 2017. The expression and significance of tyrosine hydroxylase in the brain tissue of Parkinson's disease rats. Exp Ther Med. 14(5):4813–4816. doi:10.3892/etm.2017.5124.

Chi J, Xie Q, Jia J, Liu X, Sun J, Deng Y, Yi L. 2018. Integrated analysis and identification of novel biomarkers in Parkinson's disease. Front Aging Neurosci. 10(JUN). doi:10.3389/fnagi.2018.00178.

Chung JY, Park HR, Lee SJ, Lee SH, Kim JS, Jung YS, Hwang SH, Ha NC, Seol WG, Lee J, Park BJ. 2013. Elevated TRAF2/6 expression in Parkinson's disease is caused by the loss of Parkin E3 ligase activity. Lab Investig. 93(6):663–676. doi:10.1038/labinvest.2013.60.

DeMaagd G, Philip A. 2015. Parkinson's disease and its management part 1: Disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis. P T 40(8):504–532.

Diansyah MR, Kusuma WA, Annisa. 2019. Analysis of protein-protein interaction using skyline query on Parkinson disease. In: 2019 Int Conf Adv Comput Sci Inf Syst. ICACSIS 2019. p. 175–180. doi:10.1109/ICACSIS47736.2019.8979892.

Dias V, Junn E, Mouradian MM. 2013. The Role of Oxidative Stress in Parkinson's Disease. J Parkinson's Dis. 3(4):461–491. doi:10.3233/JPD-130230.

Dorsey ER, Constantinescu R, Thompson JP, Biglan KM, Holloway RG, Kieburtz K, Marshall FJ, Ravina BM, Schifitto G, Siderowf A, Tanner CM. 2007. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68(5):384–386. doi:10.1212/01.wnl.0000247740.47667.03.

Esposito G, Ana Clara F, Verstreken P. 2012. Synaptic vesicle trafficking and Parkinson's disease. Dev Neurobiol. 72(1):134–144. doi:10.1002/dneu.20916.

Hao T, Peng W, Wang Q, Wang B, Sun J. 2016. Reconstruction and application of protein–protein interaction network. Int J Mol Sci. 17(6). doi:10.3390/ijms17060907.

Jansen R, Lan N, Qian J, Gerstein M. 2002. Integration of genomic datasets to predict protein complexes in yeast. In: J Struct Funct Genomics., volume 2. Kluwer Academic Publishers. p. 71–81. doi:10.1023/A:1020495201615.

Klein C, Westenberger A. 2012. Genetics of Parkinson's disease. Cold Spring Harb Perspect Med. 2(1). doi:10.1101/cshperspect.a008888.

Konovalova EV, Lopacheva OM, Grivennikov IA, Lebedeva OS, Dashinimaev EB, Khaspekov LG, Fedotova EY, Illarioshkin SN. 2015. Mutations in Parkinson's disease-associated PARK2 gene are accompanied by imbalance in programmed cell death systems. Acta Naturae. 7(4):146–151. doi:10.32607/20758251-2015-7-4-146-149.

Kontaki M, Papadopoulos AN, Manolopoulos Y. 2008. Continuous k-dominant skyline computation on multidimensional data streams. In: Proc ACM Symp Appl Comput. p. 956–960. doi:10.1145/1363686.1363908.

Lebouvier T, Chaumette T, Paillusson S, Duyckaerts C, Bruley Des Varannes S, Neunlist M, Derkinderen P. 2009. The second brain and Parkinson's disease. Eur J Neurosci. 30(5):735–741. doi:10.1111/j.1460-9568.2009.06873.x.

Lin X, Yuan Y, Zhang Q, Zhang Y. 2007. Selecting stars: The k most representative skyline operator. In: Proc. - Int Conf Data Eng. p. 86–95. doi:10.1109/ICDE.2007.367854.

Liu W, Wu A, Pellegrini M, Wang X. 2015. Integrative analysis of human protein, function and disease networks. Sci Rep. 5. doi:10.1038/srep14344.

Loeffler DA, Klaver AC, Coffey MP, Aasly JO, LeWitt PA. 2016. Age-related decrease in heat shock 70-kDa protein 8 in cerebrospinal fluid is associated with increased oxidative stress. Front Aging Neurosci. 8(JUN). doi:10.3389/fnagi.2016.00178.

Markaki I, Bergström S, Tsitsi P, Remnestål J, Månberg A, Hertz E, Paslawski W, Sorjonen K, Uhlén M, Mangone G, Carvalho S, Rascol O, Meissner WG, Magnin E, Wüllner U, Corvol JC, Nilsson P, Svenningsson P. 2020. Cerebrospinal Fluid Levels of Kininogen-1 Indicate Early Cognitive Impairment in Parkinson's Disease. Mov Disord. 35(11):2101–2106. doi:10.1002/mds.28192.

Mata IF, Shi M, Agarwal P, Chung KA, Edwards KL, Factor SA, Galasko DR, Ginghina C, Griffith A, Higgins DS, Kay DM, Kim H, Leverenz JB, Quinn JF, Roberts JW, Samii A, Snapinn KW, Tsuang DW, Yearout D, Zhang J, Payami H, Zabetian CP. 2010. SNCA variant associated with Parkinson disease and plasma α-synuclein level. Arch Neurol. 67(11):1350–1356. doi:10.1001/archneurol.2010.279.

Odagaki Y, Toyoshima R. 2006. Dopamine D2 receptor-mediated G protein activation assessed by agonist-stimulated [35S]guanosine 5′-O-(γ-thiotriphosphate) binding in rat striatal membranes. Prog Neuro-Psychopharmacol Biol Psychiatry. 30(7):1304–1312. doi:10.1016/j.pnpbp.2006.05.007.

Raman K, Damaraju N, Joshi GK. 2014. The organisational structure of protein networks: revisiting the centrality–lethality hypothesis. Syst Synth Biol. 8:73–81. doi:10.1007/s11693-013-9123-5.

Scardoni G, Lau C. 2012. Centralities Based Analysis of Complex Networks. In: New Frontiers in Graph Theory, chapter 16. Rijeka: IntechOpen. doi:10.5772/35846.

Scardoni G, Petterlini M, Laudanna C. 2009. Analyzing biological network parameters with CentiScaPe. Bioinformatics 25(21):2857–2859. doi:10.1093/bioinformatics/btp517.

Sharma P, Bhattacharyya DK, Kalita JK. 2016. Centrality analysis in PPI networks. In: 2016 Int Conf Access to Digit World, ICADW 2016 - Proc. p. 135–140. doi:10.1109/ICADW.2016.7942528.

Siddiqui IJ, Pervaiz N, Abbasi AA. 2016. The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication. Sci Rep. 6. doi:10.1038/srep24475.

Stumpf MPH, Thorne T, Silva dE, Stewart R, An HJ, Lappe M. 2005. Estimating thesize of the human interactome. In: Proceedings of the National Academy of Sciences. p. 6959–6964. doi:10.1073/pnas.0708078105.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Von Mering C. 2018. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47(D1):D607–D613. doi:10.1093/nar/gky1131.

Szybinska A, Lesniak W. 2017. P53 dysfunction in neurodegenerative diseases - The cause or effect of pathological changes? Aging Dis. 8(4). doi:10.14336/AD.2016.1120.

Tan JM, Wong ES, Lim KL. 2009. Protein misfolding and aggregation in Parkinson's disease. Antioxidants Redox Signal. 11(9):2119–2134. doi:10.1089/ars.2009.2490.

Trist BG, Hare DJ, Double KL. 2018. A Proposed Mechanism for Neurodegeneration in Movement Disorders Characterized by Metal Dyshomeostasis and Oxidative Stress. Cell Chem Biol. 25(7):807–816. doi:10.1016/j.chembiol.2018.05.004.

Tsolakidou A, Czibere L, Pütz B, Trümbach D, Panhuysen M, Deussing JM, Wurst W, Sillaber I, Landgraf R, Holsboer F, Rein T. 2010. Gene expression profiling in the stress control brain region hypothalamic paraventricular nucleus reveals a novel gene network including Amyloid beta Precursor Protein. BMC Genomics 11(1). doi:10.1186/1471-2164-11-546.

Twelves D, Perkins KS, Counsell C. 2003. Systematic review of incidence studies of Parkinson's disease. Mov Disord. 18(1):19–31. doi:10.1002/mds.10305.

Usman MS, Kusuma WA, Afendi FM, Heryanto R. 2019. Identification of Significant Proteins Associated with Diabetes Mellitus Using Network Analysis of Protein-Protein Interactions. Comput Eng Appl J. 8(1):41–52. doi:10.18495/comengapp.v8i1.283.

Veeriah S, Taylor BS, Meng S, Fang F, Yilmaz E, Vivanco I, Janakiraman M, Schultz N, Hanrahan AJ, Pao W, Ladanyi M, Sander C, Heguy A, Holland EC, Paty PB, Mischel PS, Liau L, Cloughesy TF, Mellinghoff IK, Solit DB, Chan TA. 2010. Somatic mutations of the Parkinson's disease-associated gene PARK2 in glioblastoma and other human malignancies. Nat Genet. 42(1):77–82. doi:10.1038/ng.491.

WHO. 2004. Atlas : Country Resources for Neurological Disorders. World Health Organization. URL https://apps.who.int/iris/handle/10665/43075.

Xiromerisiou G, Hadjigeorgiou GM, Papadimitriou A, Katsarogiannis E, Gourbali V, Singleton AB. 2008. Association between AKT1 gene and Parkinson's disease: A protective haplotype. Neurosci Lett. 436(2):232–234. doi:10.1016/j.neulet.2008.03.026.

Yu J, Fotouhi F. 2006. Computational Approaches for Predicting Protein–Protein Interactions: A Survey. J Med Sys. 30(1):39–44. doi:10.1007/s10916-006-7402-3.