DIseAse MOdule Detection (DIAMOND) algorithm in Parkinson Disease

Luca Farinola

Student ID i6326662

Course: MSB1014, Network Biology  
Program: Master Systems Biology

Faculty: FSE

Academic Year: 2023/24

**Date: 27/10/2023**

1 Introduction

Parkinson's disease (PD) affects millions worldwide, causing progressive motor and various non-motor symptom. While the exact cause is unknown, both genetic and environmental factors are believed to contribute (1). Lots of effort has been directed into studying the molecular biology behind PD using omics data analysis and next generation sequencing technologies. At the same time researchers in the field of molecular biology have curated information from various sources to create network databases that represent the interactome (2) (set of molecular interactions) of different organisms or cellular systems to study diseases. Human interactome can be extremely useful to study how genes are connected, which can be interesting information. Changes in expression levels as measured in log fold change (very common metric in bioinformatics together with statistical significance) can be very relevant but is not necessarily the only parameter we need to take into consideration when analysing a biological system. In fact, connectivity can also be very interesting to investigate (3). In this project the aim is to investigate PD using a network biology approach. The research question would be: Can we use network diffusion to explore biological processes  starting from restricted prior knowledge ?

2. Materials and Methods

**Diamond**

To perform network analysis Disease Module Detection (DIAMOnD) (4) was utilized. This is a network propagation algorithm able to expand on a graph starting from a relatively few numbers of seed nodes and select highly interacting neighbours. The algorithm was specifically developed to find diseases modules in the context of protein-protein interaction (PPI) networks. The goal of the algorithm is to identify communities of functionally related proteins and select those that have a significant number of interactions within those communities. The algorithm will perform different iterations. For each cycle it assigns to candidates nodes (having at least one link with the set of seed proteins) a p-value. Those proteins are therefore ranked and those with higher significance will be added to the set of seeds. After several iterations the result is the set of highly interacting genes. The original algorithm was written in Python, the R version of diamond used in this case can be found at the linked GitHub page (https://github.com/luca-farinola/DIseAse-MOdule-Detection-DIAMOND-algorithm-in-Parkinson-Disease).

**Data**

As input the algorithm requires a network and a set of seed proteins. Human Interactome, a graph of interactions among genes, is the backbone on which the propagation is carried out. Such network is not specific to any phenotype or biological condition and contains a huge number of links. Those are

labelled according to what kind of interactions they represent or, more precisely, the source of information. All interaction exclusively based on literature were removed while other types of links were selected to zoom on different part of the network obtained with genes selected from the algorithm. Seed nodes were chosen among the genes related to Parkinson disease in MedGen (MedGen UID: 10590). Once PD associated genes are retrieved through DIAMOnD

those are visualized and mapped to fold-change values of differentially expressed genes retrieved from the expression ATLAS reporting the result of 1 colour mRNA microarray Transcriptional analysis of prefrontal area 9 in PD (E-GEOD-20168 on array design A-AFFY-33) (5).

**Analysis and network visualization**

Diamond is applied both performing 50 and 500 iterations. Once Diamond genes are retrieved the analysis consist in observing how those genes interact. In order to do so the interactome is filtered and only seeds and diamonds genes are kept. Network visualization is performed with Cytoskape (6)(version 3 9.1) and this process automated using RCy3 (7) (api version = v1 ) that allows communication between R and Cytoskape via its REST API . Depending on the situation further filtering can be applied according to the sources (type of interaction) as discussed earlier. All created networks are reported in the saved .cys session also reported in the Result folder of the GitHub directory. Layout for better visualization is applied manually using the app graphical user interface. Lastly to have a better grasp on biological processes captured when selecting those genes enrichment analysis is performed by using ClusterProfiler (v = 4.6.2) (8).

3. Results

Each Iteration correspond to a new gene added to the set of seeds. This means that, considering the starting point of 22 seeds, the set of relevant genes with 50 iterations was enlarged to 72. The resulting network contains 175 links resulting in a very dense graph (fig1 A). Seeds protein (having an ellipse shape) seem to occupy mainly peripheral positions. In contrast nodes resulting from diamond calculations (for which the choice for the shape was obvious : diamond) are more central and tend to have in average more interactions, as expected. Nodes are coloured according to the fold change results from the expression ATLAS. A range of (-1, 1) for a continuous type of colour mapping led to a very ‘light’ colouring for most of the mapped nodes. In addition to that not all nodes were mapped to their relative fold change value resulting in yellow, set as default colour. In terms of A group of colorful dots and lines

Description automatically generatedA diagram of a network

Description automatically generateddownregulation, GSK3B and SINCA appear to be the most intriguing, with log fold changes of -1 followed by CDK5 and RHOB with a fold change of -0.6. On the other side, for upregulation FOXO1 seem to be the only gene with a relatively high fold change of 0.6. By checking the significance values from the gene expression results only GSK3B, SINCA (downregulated) and NFATC3 (slightly upregulated) have a low p-value (around 0.05 or lower). Gene Ontology on this set of genes show increase activity in protein serine/threonine kinase activity and protein threonine activity.

Fig1 : gene ontology of diamond genes after 50 iterations and their interaction

A diagram of a graph

Description automatically generated

Fig2 : Diamond after 500 iteration, links representing respectively only regulatory, signaling, complexes interactions

Diamond findings are also evaluated after 500 iterations. The resulting network was substantially bigger and had high number of edges, 4559 edges for 517 nodes thus making any visualization analysis impossible without further filtering. As mentioned already In the previous section the filtering was done for the sources (type o interaction). In the saved cytoskape session it is possible to see

different networks obtained by applying different filters. Regulatory, Signalling, Complexes and Kinase where further explored. Gene Ontologies reflect the kind of selection carried out in each case, for example regulation of miRNA for genes linked together as Regulatory, regulation of binding for genes connected through link labelled as Complexes. In coherence with what was found for the GO of the first 50 iteration the network extracted as Kinase is the largest, for that reason in that case ontology analysis is performed for both up and down regulated.

**4. Discussion**

The main goal is to explore different way to extract genes of interests using network. This report shows how network propagation can be an effective approach and can be a complementary to differential gene expression keeping a slightly different focus. Results from gene expression are indeed not perfectly matching protein extracted from network diffusion but there are still some interesting biological observations that can be done. For instance, GSK3B is known to be critical for parkinsonian degeneration in dopaminergic neurons (9), which is coherent with what was observed both the network after 50 iteration (fig) and in the subnetworks created after 500 iterations (not shown). Finding protein kinases activity among the enriched ontologies is also very coherent with what we know on PD (10). In addition, the idea of using label on interactions as filters can allow research to zoom on specific biological processes. Diamond can be a very powerful tool to put findings in molecular biology in a broader context allowing researcher to explore the intricate world o protein

interactions. In addition, the fact that the Algorithm is based on graph theory ads new levels of information as interactions among genes are not always taken into consideration in standard differential expression. To finally come back to our research question, these results show clearly how we can, starting from restricted prior knowledge (seed genes), to explore even a complex biological condition as PD.

**5. References**

1. Balestrino R, Schapira AHV. Parkinson disease. Eur J Neurol. 2020 Jan;27(1):27–42.

2. Vidal M, Cusick ME, Barabási AL. Interactome Networks and Human Disease. Cell. 2011 Mar;144(6):986–98.

3. Farina L. Network as a language for precision medicine. Ann Ist Super Sanita. 2021;57(4):330–42.

4. Ghiassian SD, Menche J, Barabási AL. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. Rzhetsky A, editor. PLOS Comput Biol. 2015 Apr 8;11(4):e1004120.

5. Zhang Y, James M, Middleton FA, Davis RL. Transcriptional analysis of multiple brain regions in Parkinson’s disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet. 2005 Aug 5;137B(1):5–16.

6. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003 Nov;13(11):2498–504.

7. Gustavsen JA, Pai S, Isserlin R, Demchak B, Pico AR. RCy3: Network biology using Cytoscape from within R. F1000Research. 2019 Dec 4;8:1774.

8. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021 Aug;2(3):100141.

9. Li J, Ma S, Chen J, Hu K, Li Y, Zhang Z, et al. GSK-3β Contributes to Parkinsonian Dopaminergic Neuron Death: Evidence From Conditional Knockout Mice and Tideglusib. Front Mol Neurosci. 2020 Jun 3;13:81.

10. Mehdi S, Rosas-Hernandez H, Cuevas E, Lantz S, Barger S, Sarkar S, et al. Protein Kinases and Parkinson’s Disease. Int J Mol Sci. 2016 Sep 20;17(9):1585.