

# Journal of Visual Language and Computing

journal homepage: [www.ksiresearch.org/jvlc](http://www.ksiresearch.org/jvlc)

## PADD: Dynamic Distance-Graph based on Similarity Measures for GO Terms Visualization of Alzheimer and Parkinson diseases

Alessia Auriemma Citarella<sup>a,\*</sup>, Fabiola De Marco<sup>a</sup>, Luigi Di Biasi<sup>a</sup>, Michele Risi<sup>a</sup> and Genoveffa Tortora<sup>a</sup>

<sup>a</sup>Department of Computer Science, University of Salerno, 84084 Fisciano (SA), Italy

### ARTICLE INFO

#### Article History:

Submitted 7.31.2021

Revised 8.11.2021

Accepted 8.20.2021

#### Keywords:

Protein Visualization

Gene Ontology

Clustering

### ABSTRACT

In the biological field, having a visual and interactive representation of data is useful, particularly when there is a need to investigate a large amount of multilevel data. It is advantageous to communicate this knowledge intuitively because it helps the users to perceive the dynamic structure in which the correct connections are present and can be extrapolated. In this work, we propose a human-interaction system to view similarity data based on the functions of the *Gene Ontology* (Cellular Component, Molecular Function, and Biological Process) of the proteins/genes for Alzheimer disease and Parkinson disease. The similarity data was built with the Lin and Wang measures for all three areas of Gene Ontology. We clustered data with the K-means algorithm in order to demonstrate how information derived from data can only be partial when using traditional display methods. Then, we have suggested a dynamic and interactive view based on SigmaJS with the aim of allowing customization in the interactive mode of the analysis workflow by users. To this aim, we have developed a first prototype to obtain a more immediate visualization to capture the most relevant information within the three vocabularies of Gene Ontology. This facilitates the creation of an omic view and the ability to perform a multilevel analysis with more details which is much more valuable for the understanding of knowledge by the end users.

© 2021 KSI Research

## 1. Introduction

In the latest years, it is becoming increasingly vital to have an omic vision in order to define biological systems at an ever-increasingly granular level. The goal of omic sciences is to generate useful knowledge which can be utilized to feature and interpret biological systems [18].


For omic sciences we refer to the wide range of biomolecular disciplines characterized by the suffix -omics including genomics, transcriptomics, proteomics, and metabolomics. In this perspective, technological innovation aids the growth of complex system biology by allowing researchers to investigate various intrinsic and extrinsic influences and events

at the base of life. Biological data is multidimensional and highly interdependent. The current challenge is to acquire a more detailed integrative view of the dynamics of cellular processes in a cell or organism enriched in biological and spatial-temporal information [19]. For this purpose, clear visualization methods can provide more immediate access to their content information.

The visualization of biological data has become increasingly relevant in Biosciences, as O'Donoghue *et al.* [14] point out because it helps researchers to interpret heterogeneous data more quickly and easily. One of the most current issues in omic data analysis is the inability to investigate relationships between multi-omic states to incorporate them and combine higher-level expertise [23].

In this paper, we report the preliminary results achieved regards visualization of the similarity of the proteins based on the protein annotations. Protein similarity visualization not based on sequence alignment can be tricky due to inter-class dissimilarities and inter-class similarity [1]. Clustering

\*Corresponding author

 [aauriemmaitarella@unisa.it](mailto:aauriemmaitarella@unisa.it) (A. Auriemma Citarella);

[fdemarco@unisa.it](mailto:fdemarco@unisa.it) (F. De Marco); [ldibiasi@unisa.it](mailto:ldibiasi@unisa.it) (L. Di Biasi);

[mrisi@unisa.it](mailto:mrisi@unisa.it) (M. Risi); [tortora@unisa.it](mailto:tortora@unisa.it) (G. Tortora)

ORCID(s): [0000-0002-6525-0217](https://orcid.org/0000-0002-6525-0217) (A. Auriemma Citarella);

[0000-0003-4285-9502](https://orcid.org/0000-0003-4285-9502) (F. De Marco); [0000-0002-9583-6681](https://orcid.org/0000-0002-9583-6681) (L. Di Biasi);

[0000-0003-1114-3480](https://orcid.org/0000-0003-1114-3480) (M. Risi); [0000-0003-4765-8371](https://orcid.org/0000-0003-4765-8371) (G. Tortora)

DOI reference number: 10.18293/JVLC2021-N1-013

and Machine Learning methods may not be able to extract interdependencies between objects effectively [9]. This fact often does not allow us to generate a clear visual representation of the information.

Our goal is to demonstrate how a human-assisted dynamic graph construction can help abstract functional relationships between proteins in order to generate a clear data visualization when a traditional clustering technique fails. For this contribution, we focused on two diseases: *Alzheimer* and *Parkinson*, the two most common neurodegenerative conditions. Alzheimer's disease (AD) is a form of degenerative dementia that occurs after 65 years. In this pathology, there is a deposition of an  $A\beta$  peptide B with the formation of senile plaques and the intracellular aggregation of *tau* protein [5]. Parkinson's disease (PD) is the second most common neurodegenerative disorder in the senile age in which neuronal loss is found in the substance nigra and formation of  $\alpha$ -synuclein aggregates that are neuropathological [15].

These pathologies show similar neurodegeneration mechanisms supported by scientific evidence with genetic, biochemical, and molecular studies. Pathological pathways involving  $\alpha$ -synuclein and *tau* proteins, oxidative stress, mitochondrial dysfunction, iron pathway, and *locus coeruleus* are among these findings [22]. Because of the overlap in their pathogenic mechanisms, they were chosen as an example for our search workflow. This feature introduces intra- and extra-class overlaps which can deceive typical clustering algorithms.

This paper is an extension of the work *Gene Ontology Terms Visualization with Dynamic Distance-Graph and Similarity Measures* [2]. We have restructured some sections of the paper, enriching the description of the approach with more details. We included two new figures (Figure 4 and Figure 5) which depict the graphical representation of the molecular function of AD and PD proteins, respectively. In addition, the chord diagrams of the recoverable information following the usage of similarity matrices have been provided as an overview (Figures 9-12). We also added further results in Section 5. In particular, we calculated the similarity between all the proteins of both diseases for the molecular function, the biological process and the cellular component. We also extracted the proteins in common to AD and PD, giving an overview of the information that can be recovered from these findings.

The paper is structured as follows. In Section 2 we describe the most important related works in the examined field. In Section 3 we discuss respectively the datasets, methodology, and performance measures which we have used in our research. Finally, we expose the visual results in Section 4 and overall results in Section 5. The conclusions with future work are outlined in Section 6.

## 2. Related Work

In the literature, several web interfaces can query the terms of the Gene Ontology. The *Gene Ontology* (GO) is a bioinformatics project which uses ontologies to enable the standardization of biological information regarding gene and

gene products properties. It is structured as an acyclic oriented graph where each GO-term is identified by a word or strings and a unique alphanumeric code [8]. The GO database is the most widely utilized resource for enrichment analysis.

*QuickGO* allows us to find and display GO terms and generate a list of correspondence results based on the user's question. This tool returns a directed acyclic graph (DAG) containing a single GO term and its associated terms and annotations. It is designed with JavaScript, Ajax, and HTML. Statistics with interactive graphs and views of term location tables are available on the fly, indicating which words are frequently noted simultaneously. The user can create a subset of annotations based on different parameters (Specific protein, Evidence Codes, Qualifier Data, Taxonomic Data, Go Terms) and download them [3].

*Gorilla*<sup>1</sup> identifies enriched GO terms in ordered lists of genes using simple, intuitive, and informative graphics, without explicitly requiring the user to provide targets or background sets. It is a GO analysis tool that employs a statistical approach with flexible thresholds to identify GO terms significantly enriched at the top of a classified gene (very useful when genomic data can be represented as a classified list of genes). The analysis's results are presented in the form of a hierarchical structure that allows for a clear view of the GO terms [6].

*Blast2GO* (B2G)<sup>2</sup> is an interactive platform that supports non-model species functional genomic research. It is a data sequence-based tool that combines high-performance analysis techniques and evaluation statistics with a high degree of user interaction. Similarity searches produce results on direct acyclic graphs [4].

*NaviGO*<sup>3</sup>, in order to measure the similarity or relation between the terms of the GO, use six different scores: Resnik, Lin, and the relevant semantic Similarity score for semantic similarity, and *Co-occurrence Association Score* (CAS), *PubMed Association Score* (PAS), and *Interaction Association Score* (IAS) for GO associations. A *Funsim* score for functional similarity is also introduced [21].

More recently, the open-source software *AEGIS* allows us to visually explore the GO data in real-time, taking into input the entire dataset GO. Any Go terms can be chosen as the anchor and have a root, leaf, or waypoint, represented with a DAG. Each source can include all the descendants of the anchor term, the leaves will only include the ancestors, and the Waypoint anchors will constitute a DAG consisted of both ancestors and descendants [25].

## 3. Methods

In this work, we have used the R environment<sup>4</sup>, a free software environment for statistical computing and graphics, and SigmaJS, a JavaScript library dedicated to graph draw-

<sup>1</sup>Gorilla: <http://cbl-gorilla.cs.technion.ac.il>

<sup>2</sup>Blast2GO: <https://www.biobam.com>

<sup>3</sup>NaviGO: <https://kiharalab.org/web/navigo/views/goset.php>

<sup>4</sup>R: <https://www.r-project.org>

ing<sup>5</sup>. We used the standard SigmaJS renderer to show the graph view.

### 3.1. Datasets

Protein datasets for AD and PD, belonging to *Homo Sapiens*, were downloaded from UNIPROT [17]. Data cleaning has been carried out, removing all duplicates. Furthermore, for each UNIPROT ID, the reference gene has been obtained and linked to the STRING. STRING database allows us to consider any protein-protein interactions (PPI) based on a score calculated on experimental evidences [16]. This step is required to eliminate those proteins that are not mapped in the database and do not have the protein-protein interaction that we are looking for. We have recovered a total of 216 genes for AD and 137 genes for PD.

### 3.2. Gene Ontology

The Gene Ontology is based on two types of relationships between objects: *instances* and *part of*. All organisms share three biological domains which can be considered as structured and controlled vocabularies:

- *Biological Process*: refers to all those events that take place within an organism resulting from an orderly set of molecular functions;
- *Cellular Component*: concerns the location of the entity in question at the level of cellular and/or subcellular structures;
- *Molecular Function*: describes the processes that occur at the molecular level.

We have identified these domains as biological process (BP), cellular component (CC), and molecular function (MF). We have recovered from UNIPROT<sup>6</sup> all the GO terms belonging to these three fields both for Alzheimer’s and Parkinson’s diseases with UniProt package in R.

### 3.3. Experimental setup

We explored two ways to calculate semantic similarity. In the first case, we calculated the similarity between proteins of Alzheimer disease and proteins of Parkinson disease for all three ontology gene domains. We considered both Lin’s similarities and Wang’s method. For simplicity, in this work we only show the results concerning the similarity of Lin while the future tool will allow user the setting of both measures. Subsequently, we clustered the data obtained for both similarity measures in BPs, CCs, and MFs domains for AD and PD with the K-means algorithm, trying with  $n=3$  and  $n=5$  clusters. In the second case, we calculated the similarity based on the Wang and Lin methods between the two sets of protein data of diseases about BPs, DCs and MFs domains in order to compare these measures.

<sup>5</sup>SigmaJS: <https://sigmaj.s.org>

<sup>6</sup>UniProt: <https://www.uniprot.org>

### 3.4. Distance Metrics

We used two types of metric to compute pairwise semantic similarities, *Lin* and *Wang*, calculated with the GOSemSim package in R [24].

#### 3.4.1. Lin’s measure

Lin measure is based on *information content* (IC). The negative log of a concept’s probability is formally known as IC. This method computes the ratio between the amount of “common information” and the amount of “total information” in the descriptions regards an object pair. This ratio corresponds to the similarity between two objects [12].

In this case, this approach can measure the similarity of the knowledge content of the GO terms for each protein dataset, proteins of AD e proteins of PD. The frequency of two GO words and their closest common ancestor in a particular corpus of GO annotations are used in the estimation. The term *Least Common Subsummer* (LCS) suggests the most basic definition that two concepts share as an ancestor. So, we can consider the following Equation 1:

$$sim_{lin} = \frac{2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (1)$$

where  $c_1$  and  $c_2$  are two concepts,  $IC$  is the information content and  $lcs$  is the function that computes the least common subsammer. In our experiment,  $c_1$  and  $c_2$  reflect the concepts represented by the GO terms referring to the BP, CC, and MF domains. The similarity is calculated for both AD and PD across all proteins in the pathological reference dataset.

#### 3.4.2. Wang measure

The Wang method is based on a *graph-based* semantic similarity. The GO terms are converted into a numeric value by aggregating the terms of their ancestors in a GO graph [20].

Given two GO terms,  $A$  and  $B$ , we can represent  $DAG_A = (A, T_A, E_A)$  and  $DAG_B = (B, T_B, E_B)$ , where  $T_n$  is the set of GO terms including the term  $n$  and all of its ancestor terms in the GO graph while  $E_n$  are the semantic relations represented as edges between the GO terms. The semantic similarity between these two terms are calculated as in Equation 2:

$$S_{GO}(A, B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)} \quad (2)$$

where  $S_A(t)$  and  $S_B(t)$  denote the S-value of a GO term  $t$  related to term  $A$  and term  $B$ .

Wang measures the semantic meaning of GO term  $n$ ,  $SV(n)$ , after obtaining the S-values for all terms in  $DAG_n$  with the Equation 3, represented below:

$$SV(n) = \sum_{t \in T_n} S_n(t) \quad (3)$$

### 3.5. K-means

K-means is one of the most common and widely used partitioning clustering algorithms which divides a set of objects into K clusters based on their attributes [13]. A cluster

is simply an aggregation of data based on similarities. The division into  $K$  clusters is done *a priori*, based on the goal to be achieved or using heuristic techniques and the clusters represent the number of centroids required by the dataset. A centroid is a real or imaginary point that represents the center of the cluster and it is updated with each algorithm iteration.

The procedure is composed by four steps:

- *Step 1:* determine the value of  $K$ ;
- *Step 2:* randomly select  $K$  points as initial centers of the clusters;
- *Step 3:* assign each new point to the cluster with the closest Euclidean distance to its center. Formally, if  $c_i$  is a centroid of the set of centroids  $C$  then each point  $x$  will be assigned to a cluster based on the following equation (Equation 4):

$$\arg \min_{c_i \in C} \text{dist}(c_i, x)^2 \quad (4)$$

where  $\text{dist}(\cdot)$  represents the Euclidean distance;

- *Step 4:* recalculate the updated cluster centers by averaging the points associated with each cluster (Equation 5):

$$c_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j \quad (5)$$

where  $S_i$  is the cluster's set of points.

The procedure repeats steps 3 and 4 until a convergence is achieved. The algorithm ensures speed of execution while leaving the data free to group and move away. For the purpose of our study, the maximum number of clusters of the K-means is limited to five. No PCA techniques were used. When the concept of similarity associated to the GO is considered, this constraint is tied to the core premise that a smaller number of clusters can be useful for biological scope. At the same time, when  $K$  is less, the K-means allows us to preserve this information but not to view it intuitively. Without a clear display of the data, the end user could not correctly interpret the results. It is necessary to represent such data as clearly as possible in order to translate it into knowledge. We attempted to collect the various forms of information from the three GO domains in order to organize and view them together.

### 3.6. Dynamic Distance-Graph

Based on the information presented in the previous sections, we propose a *dynamic build cyclic distance graph* (DCDG) to visualize and transfer knowledge regarding the GO terms. Our goal is to provide a clearer visualization of the GO interconnections than other visualization methods like clustering or partitioning. We used a web-based workspace built with Javascript and SigmaJS to allow the user to explore this interconnection. Workspace is designed to be as clean as possible. It starts as an empty web app with

a single callable overlay menu on the upper left corner, allowing users to search the entry point protein into datasets.

The BP, CC, and MF distance matrices, calculated before the execution of the k-means algorithm, were used as input datasets. When selected, the entry protein becomes the root of the graph. Users can click on each graph node to show a context menu (as depicted in Figure 1) in which it is possible to choose extension (explosion) operation for the node itself.

We defined three kinds of extensions for this contribution, each of them related to one dataset: BP, CC and MF, whose definitions are those intended by the three vocabularies of the GO. The distance between each node pairs is written on the arcs between them. This value, which defines the similarity measure, provides the reading key to display protein through the dynamic build cyclic distance graph. Proteins are connected to each other from these values that allow us to explore the graph taking into account the resemblance values between biological process, molecular function and cellular component. Also, the distance value is used to separate nodes into spaces.

The ForceAtlas2 algorithm is used to avoid overlapping between near nodes. In particular, we used ForceAtlas2 embedded into SigmaJS [11]. ForceAtlas2 is a layout algorithm for force-directed graphs. This algorithm allows us to position each node depending on the other nodes using the distances between them as edge weights. Just because of this condition, the position of a node must always be confronted with the other nodes. The fundamental advantage of using ForceAtlas2 for the representation of protein graphs is to have an easier view of the structure because the structural proximity present in the original datasets is converted to visual proximity.

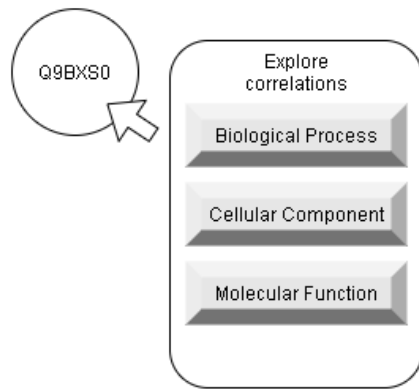
In order to better empathize the functionality distance between GO, we defined a spatial distance  $SD$  with the following equation (Equation 6). Given two nodes, A and B and their own distance  $d$ :

$$SD = \log_e(d) \quad (6)$$

where  $d$  is the distance and the  $\log_e$  is the natural logarithm with the number of Nepero as base.

Note that  $SD$  is used only for graphical purposes in the rendering routines. Figure 7 shows no linear proportionality into edge lengths: see the distance between (Q8IZY2, Q9BS0) and (Q93045, Q9BS0). Still, for graphical purposes, we defined a threshold  $th_i$  as the mean of all the distances into the dataset  $i$  used for node expansion. As an example, given the node Q9BXS0 (see Figure 7), the threshold for the protein Q9BXS0 is the mean of the edge's weight between Q9BXS0 and the related nodes. When the distance  $SD$  between two node A and B is greater than  $th_i$ , then node A and B are considered belonging to a different cluster. A dotted line renders each class separation. For the first prototype of the proposed method see the [Prototype Page](#)<sup>7</sup>. The input requires a symmetry (or distance) matrix in TSV format. After clustering, it is also possible to download the table of coordinates between the various proteins, represented here graph-

<sup>7</sup><https://smcovid19.org/simtest/>



**Figure 1:** The contextual menu is available for each node.

ically as dynamic dots. The prototype is still being updated for further improvements to guarantee the user full control of the visualization process.

## 4. Results

### 4.1. K-means visualization

Figures 2-5 report how the GO objects are partitioned regarding the BP and MF features for AD and PD, with  $K$  equal to 3 and 5. The axis reports the distance between each item to its centroid. We used `cluster` and `factoextra` packages in R to perform clusterization. We considered only the Lin's measure for graphical example. We have found that clustering with the K-means algorithm produces visually misleading and uninformative overlaps. This is due to the density of clusters that involve very close intra-cluster distances.

### 4.2. DCDG visualization

To test our methods, we used protein data based on calculated similarity of Lin. In particular, we considered the G9BXS0 protein from the similarity matrices and we identified the proteins of its neighborhood to build our view of node expansion. Before testing DCDG view, we carried out a simple statistic of the common GO terms, for the only BP component, between this *root* protein and its neighbors. We represented them with a Venn diagram [10] (see Figure 6), on the basis of GO Lin's similarity matrix.

In this scenario, each protein is represented by a closed curving line in the Venn diagram (a circle). A set of GO terms is associated with each protein. In our representation the overlapping area of the circles measures the size of common GO terms for the BP among the proteins. So, this view allows us to evaluate how many common elements are among the different sets of the terms GO for all the selected proteins. It is evident that a simple analysis of terms provides no helpful information beyond the simple observation that there are terms common to all five sets of GO terms for each protein. Instead, introduce similarity based on the *information content* of the GO terms is useful for expanding knowledge regarding biological aspects that would be omitted by a simple statistical analysis.

Figure 7 shows the BP expansion with the DCDG view for the node G9BXS0, a protein produced by *COL25A1* gene for *Homo Sapiens* organism. This protein inhibits the fibrillization of  $\beta$ -amyloid peptide which constitutes amyloid plaques present in Alzheimer's disease. It also assembles the amyloid fibrils in aggregates which are resistant to the demerger mechanisms.

The DCDG view allows the user to see and understand immediately the proteins belonging to the two distinct BP classes: **CLASS 1**, related to many biological processes such as signaling pathway and positive and negative regulation of cellular and chemical complexes and **CLASS 2**, concerning the organization of fibrils, microtubules, and structures of the cytoskeleton.

Figure 8 highlights the successive expansion of Q8IZY2 and Q9POL2 proteins. Due to distances, a new class was identified by the system (**CLASS 3**). In terms of biological meaning, the visualization clearly shows that the additional third class emphasizes further involvement of proteins indicated in different biological processes compared to previous classes. In particular, this class intervenes in broader biological regulation processes involving energy homeostasis and cell cycle regulation systems.

## 5. AD and PD similarity

Diseases similarity can be determined based on three domains of the GO: molecular function (MF) similarity, biological process similarity (BP) and cellular component similarity (CC). We used the `GoSemSim` package [24]. We used Wang's technique, which leverages the graph structure topology for the GO to compute semantic similarity between the two sets of Alzheimer's and Parkinson's proteins. We have also calculated the similarity of Lin, based on the IC of the three GO domains, between AD e PD in order to compare the differences between these two used methods, as reported in Table 1. We can note as the values are similar for both similarity measure, except for a 5% waste for BP. In Table 2 we reported the common proteins between the two diseases with their ID UNIPROT and the description for each of them. Based on the similarities of BP, MF and CC, we can build a protein network for each of the three domains under consideration. This could respond to the end user request regarding the presence of similar proteins in the function, biological process or cellular location of a series of disorders. As example, in Figure 9 and Figure 10, the similarities of the BP and MF domains for the P03886 protein, present in the AD and PD, are shown. The threshold chosen for the representation is 80%. The protein in question is highlighted in the chord graph. With the threshold previously chosen for BP and MF, the similarities between proteins in PD and AD are depicted as a whole in Figure 11 and Figure 12.

## 6. Conclusion

Graphs are the most natural way to model interactions between entities in many fields. Dynamic graph representations result from the intrinsically dynamic character of such

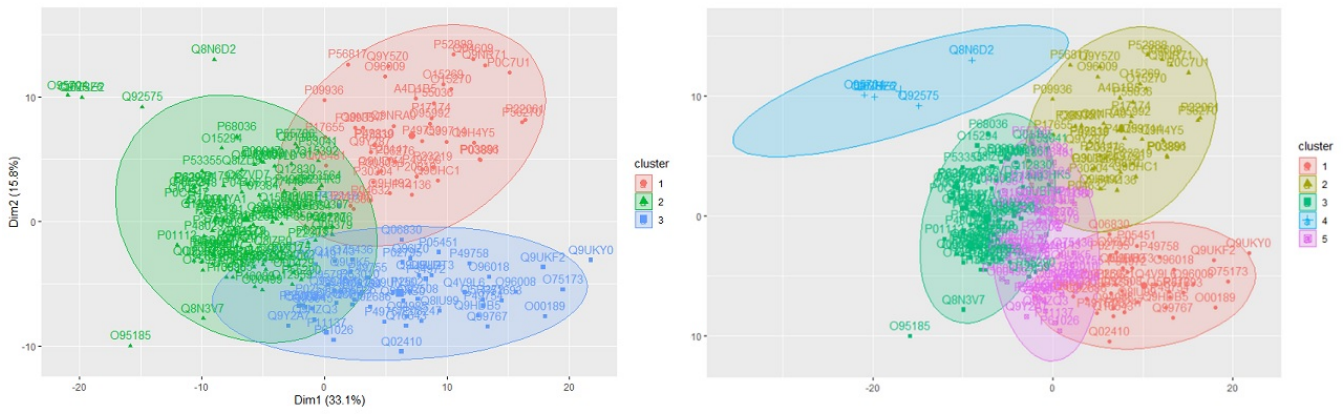


Figure 2: K-means for BP for AD with Lin's measure ( $K=3$  on the left and  $K=5$  on the right).

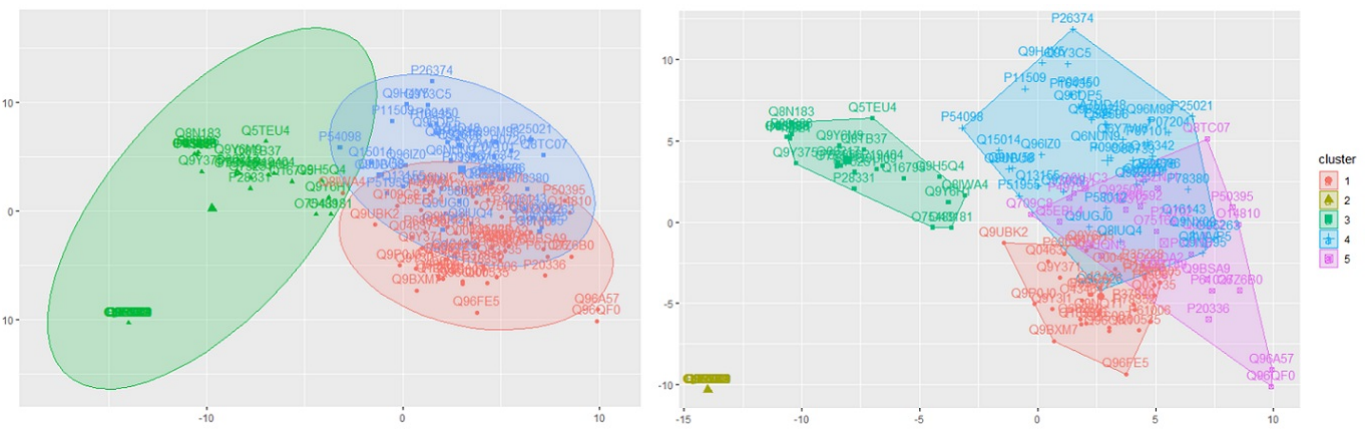


Figure 3: K-means for BP for PD with Lin's measure ( $K=3$  on the left and  $K=5$  on the right).

data [7]. In this paper, we explored an alternative way to graphically view the relationships between the GO terms based on their information content. In particular, we have proposed a *human interaction*-based viewing system that allows the users to have a complete omic vision of data. In particular, we ensured the direct representation of the inter-

class and intra-class correlations between involved proteins. The strategy proposes an instrument to investigate the GO with a customizable and flexible approach providing information to a more general or selective level.

We presented a distance cyclic distance graph (DCDG) as a GO terms visualization approach to immediately repre-

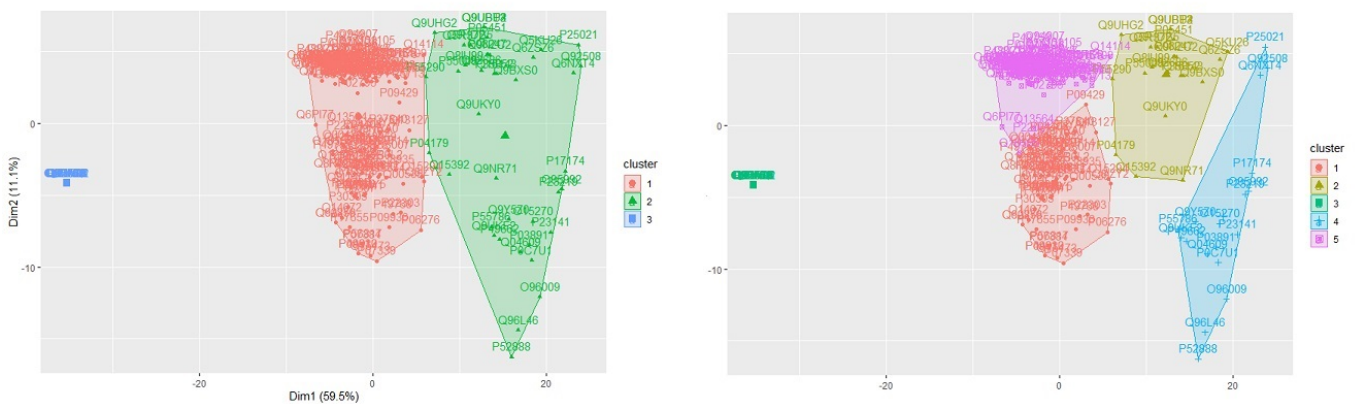


Figure 4: K-means for MF for AD with Lin's measure ( $K=3$  on the left and  $K=5$  on the right).

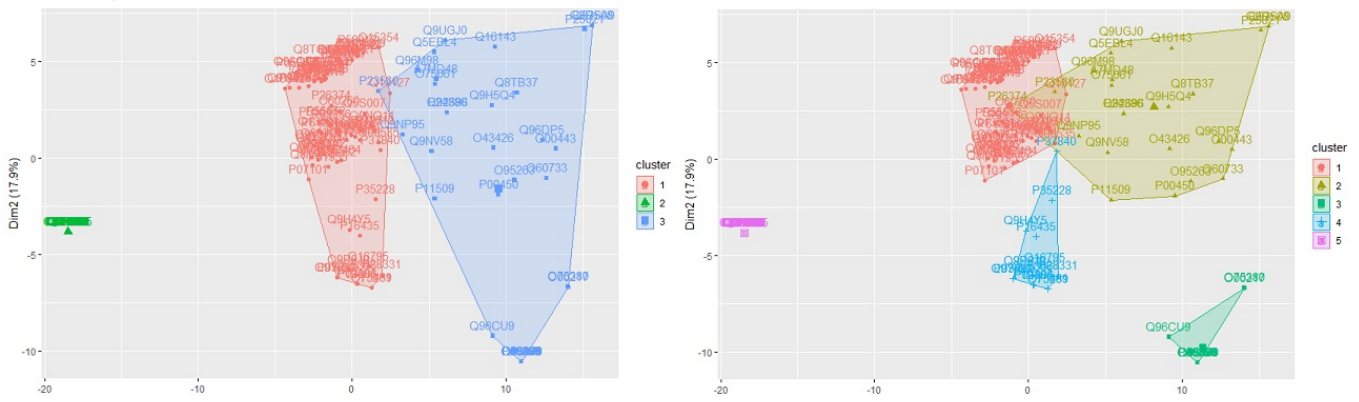


Figure 5: K-means for MF for PD with Lin's measure ( $K=3$  on the left and  $K=5$  on the right).

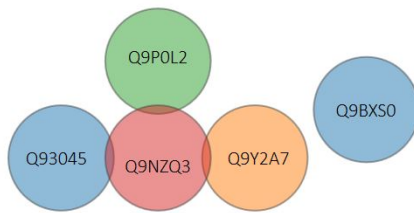


Figure 6: Venn Diagram for G9BXS0.

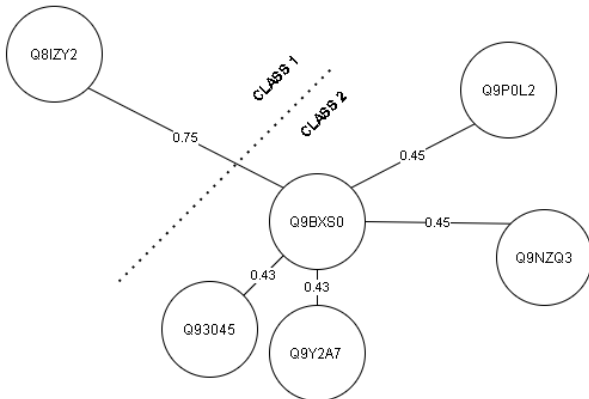


Figure 7: The result of Q9BXS0 expansion by BP dataset.

Table 1  
Similarity value for AD and PD.

Measure	BP similarity	MF similarity	CC similarity
Wang	88.3%	91.3%	96.7%
Lin	93%	92%	96.6%

sent interconnection between elements. The prototype was written as a web app by using the SigmaJS framework.

We used two similarity methods, Lin's and Wang's measure, on the three GO vocabularies (*Biological Process, Cel-*

Table 2  
Common proteins in AD and PD.

UNIPROT ID	
P03886	NADH-ubiquinone oxidoreductase chain 1
P05067	Amyloid-beta precursor protein
P09936	Ubiquitin carboxyl-terminal hydrolase isozyme L1
P10636	Microtubule-associated protein tau
P25021	Histamine H2 receptor
P37840	Alpha-synuclein
P49754	Vacuolar protein sorting-associated protein 41 homolog
P61026	Ras-related protein Rab-10
P68036	Ubiquitin-conjugating enzyme E2 L3
P78380	Oxidized low-density lipoprotein receptor 1
Q55007	Leucine-rich repeat serine/threonine-protein kinase 2
Q9H4Y5	Glutathione S-transferase omega-2
Q96I20	PRKC apoptosis WT1 regulator protein
Q00535	Cyclin-dependent-like kinase 5
Q13127	RE1-silencing transcription factor
Q13501	Sequestosome-1
Q16143	Beta-synuclein
Q92508	Piezo-type mechanosensitive ion channel component 1
Q92876	Kallikrein-6

ular Component and Molecular Function) for two neurodegenerative diseases, Alzheimer and Parkinson. Thanks to these metrics, we built three different distance matrices (BP, CC, and MF) for each condition.

We explored the differences between the standard cluster view and the proposed DCDG view. The datasets were clustered using the K-means algorithm to show a classic clustering plot. Also, we use the proposed DCDG method to plot the same information into a graph view.

By applying a classic display of clustering, visually was not possible to recover the information immediately, also due to the problem of overlapping of some clusters elements. On

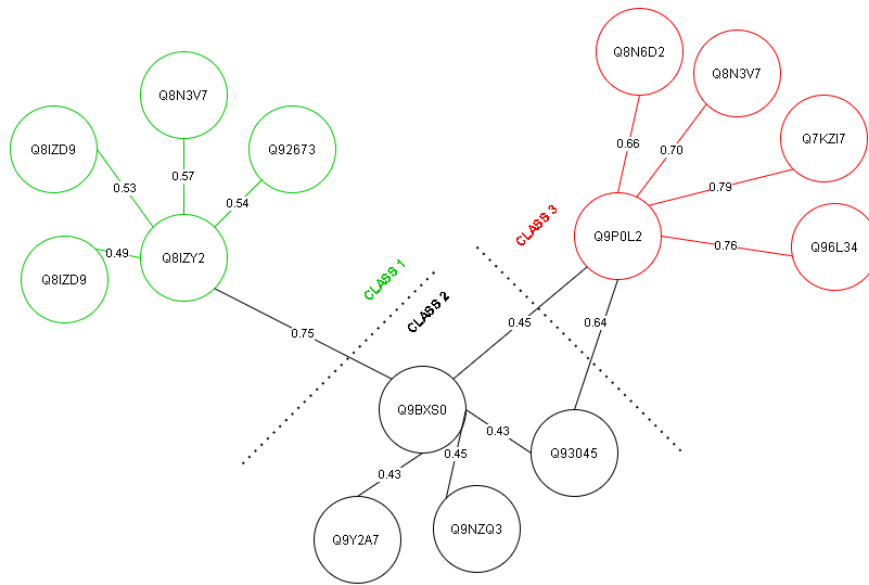


Figure 8: The result of Q8IZY2 and Q9P0L2 expansion by BP dataset.

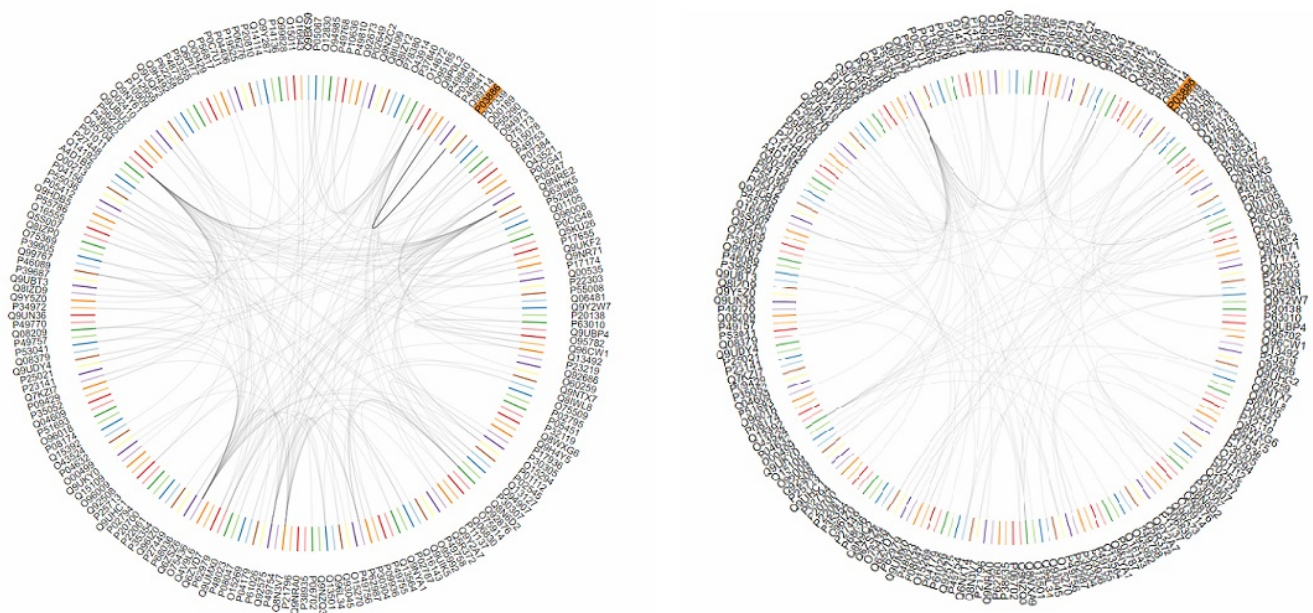


Figure 9: Similarity of BP (on left) and MF (on right) for the protein P03886 in AD.

the other hand, the display with DCDG allows a more immediate understanding of the interactions present between the proteins based on the similarity representative of the three vocabularies of the GO. The existence of well-outed protein clusters in a system is one of the purposes of our work as it represents a fundamental topological characteristic to understand the entire network of connections. This subdivision makes it possible to view the existing relationships between proteins and provides a tool which meets the need to identify and understand why some structural elements are grouped at different levels (cellular, biological and molecular) of in-

depth.

As future work, we plan to improve the web-based tool prototype into a web app with more functionality for the user for exploring protein data based on the proposed assumptions in this research study, guaranteeing user-target customization of the tools available.

## References

- [1] Arif, M., 2012. Similarity-dissimilarity plot for visualization of high dimensional data in biomedical pattern classification. *Journal of Medical Systems* 36, 1173–1181.



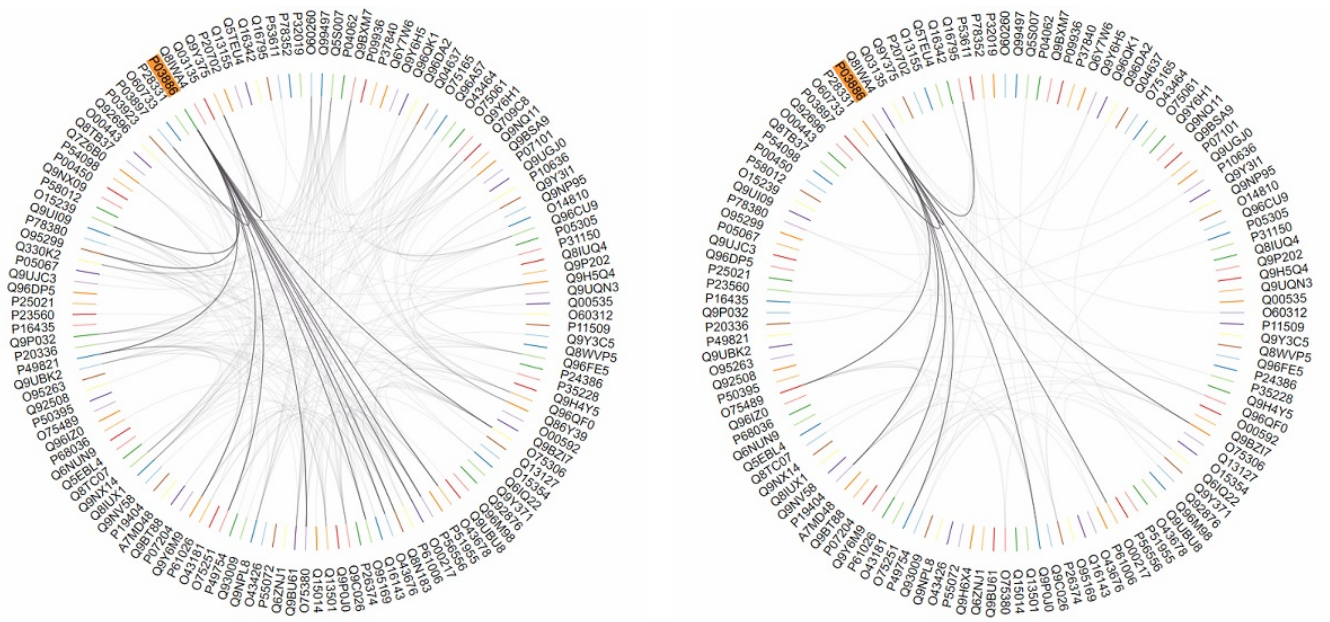


Figure 10: Similarity of BP (on left) and MF (on right) for the protein P03886 in PD.

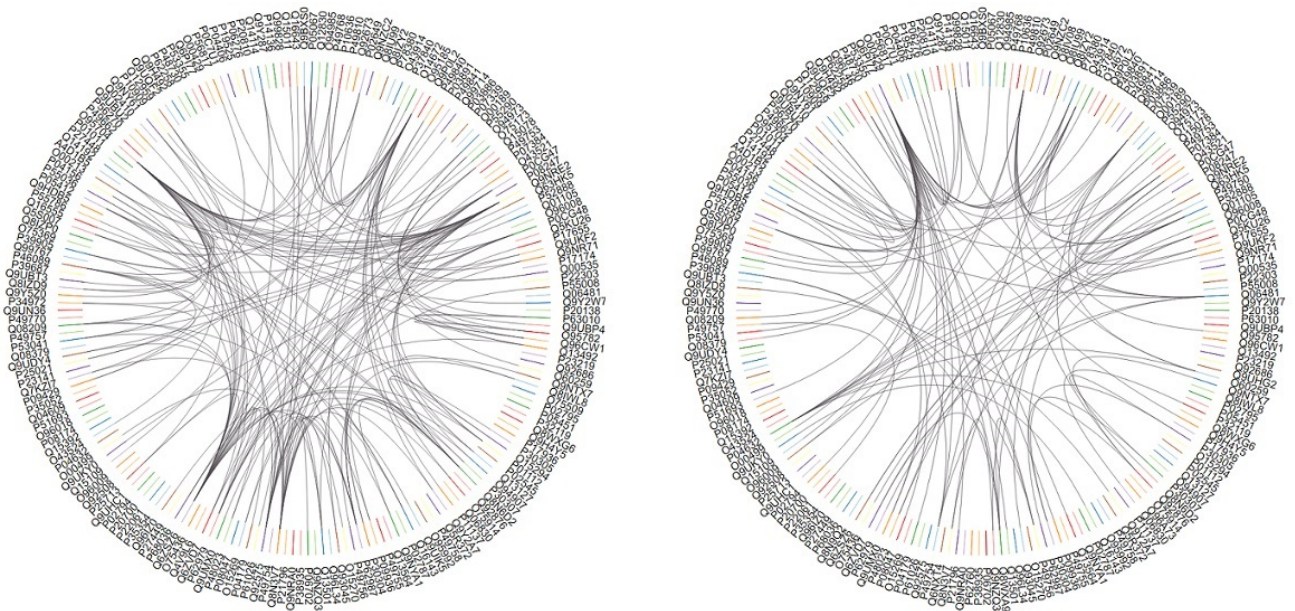


Figure 11: Similarity of BP (on left) and MF (on right) in AD.

[2] Auriemma Citarella, A., De Marco, F., Di Biasi, L., Risi, M., Tortora, G., 2021. Gene ontology terms visualization with dynamic distance-graph and similarity measures, in: 27th International Distributed Multimedia Systems Conference on Visualization and Visual Languages, DMSIVA 2021, Knowledge Systems Institute Graduate School, KSI Research Inc., pp. 85–91.

[3] Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'donovan, C., Apweiler, R., 2009. QuickGO: A web-based tool for gene ontology searching. *Bioinformatics* 25, 3045–3046.

[4] Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676.

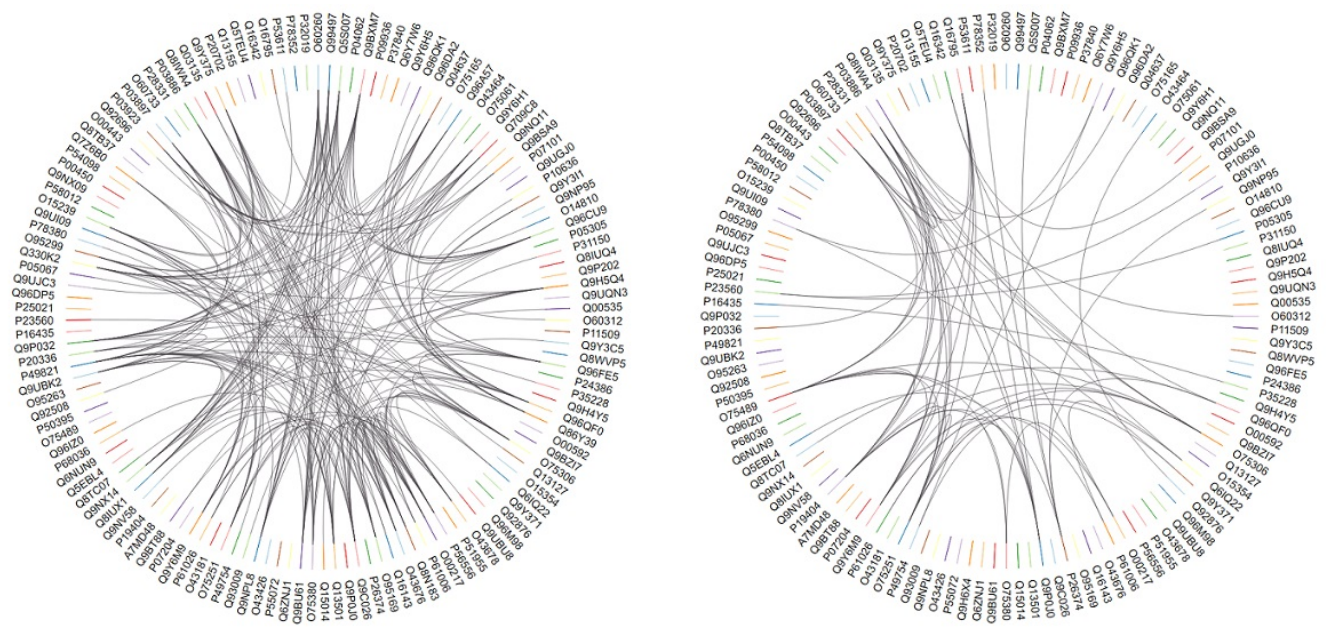
[5] Duyckaerts, C., Delatour, B., Potier, M.C., 2009. Classification and basic pathology of Alzheimer disease. *Acta Neuropathologica* 118, 5–36.

[6] Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z., 2009. GOrilla: A tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics* 10, 1–7.

[7] Fenn, D.J., Porter, M.A., Mucha, P.J., McDonald, M., Williams, S., Johnson, N.F., Jones, N.S., 2012. Dynamical clustering of exchange rates. *Quantitative Finance* 12, 1493–1520.

[8] Gene Ontology Consortium, 2008. The gene ontology project. *Nucleic Acids Research* 36, D440–D444.

[9] Goyal, M., Knackstedt, T., Yan, S., Hassanpour, S., 2020. Artificial intelligence-based image classification for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, 104065.



**Figure 12:** Similarity of BP (on left) and MF (on right) in PD.

[10] Henderson, D.W., 1963. Venn diagrams for more than four classes. *The American Mathematical Monthly* 70, 424–426.

[11] Jacomy, M., Venturini, T., Heymann, S., Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS one* 9, e98679.

[12] Lin, D., 1998. Extracting collocations from text corpora, in: *Proceedings of the First Workshop on Computational Terminology*, pp. 57–63.

[13] MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

[14] O’Donoghue, S.I., Gavin, A.C., Gehlenborg, N., Goodsell, D.S., Hériché, J.K., Nielsen, C.B., North, C., Olson, A.J., Procter, J.B., Shatuck, D.W., et al., 2010. Visualizing biological data—now and in the future. *Nature Methods* 7, S2–S4.

[15] Poewe, W., Seppi, K., Tanner, C.M., Halliday, G.M., Brundin, P., Volkman, J., Schrag, A.E., Lang, A.E., 2017. Parkinson disease. *Nature Reviews Disease Primers* 3, 1–21.

[16] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al., 2016. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, gkw937.

[17] UniProt Consortium, 2015. UniProt: A hub for protein information. *Nucleic Acids Research* 43, D204–D212.

[18] Vailati-Riboni, M., Palombo, V., Loor, J.J., 2017. What are omics sciences?, in: *Periparturient Diseases of Dairy Cows*. Springer, pp. 1–7.

[19] Veenstra, T.D., 2021. Omics in systems biology: Current progress and future outlook. *Proteomics* 21, 2000235.

[20] Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F., 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281.

[21] Wei, Q., Khan, I.K., Ding, Z., Yerneni, S., Kihara, D., 2017. NaviGO: Interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *Bmc Bioinformatics* 18, 1–13.

[22] Xie, A., Gao, J., Xu, L., Meng, D., 2014. Shared mechanisms of neurodegeneration in Alzheimer’s disease and Parkinson’s disease. *BioMed Research International*.

[23] Yan, J., Risacher, S.L., Shen, L., Saykin, A.J., 2018. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Briefings in Bioinformatics* 19, 1370–1381.

[24] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S., 2010. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26, 976–978.

[25] Zhu, J., Zhao, Q., Katsevich, E., Sabatti, C., 2019. Exploratory gene ontology analysis with interactive visualization. *Scientific Reports* 9, 1–9.