Gene Ontology Terms Visualization with Dynamic Distance-Graph and Similarity Measures

Alessia Auriemma Citarella, Fabiola De Marco, Luigi Di Biasi, Michele Risi, Genoveffa Tortora
Department of Computer Science
University of Salerno
84084 Fisciano (SA), Italy
{aauriemmacitarella, fdemarco, ldibiasi, mrisi, tortora}@unisa.it

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 see the dynamic structure in which the correct connections are interacting and 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) for Alzheimer's and Parkinson's disease proteins/genes. 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 and 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. In this way we have obtained a more immediate visualization to capture the most relevant information within the three vocabularies of Gene Ontology. This facilitates to obtain an omic view and the possibility of carrying out a multilevel analysis with more details which is much more useful in order to better understand the knowledge of the end user.*

Index terms— Protein Visualization, Gene Ontology, Clustering

1 Introduction

The importance of having an omic vision is becoming increasingly important to define biological systems at an increasingly detailed level. Omic sciences aim to produce useful knowledge that can be used to characterize and interpret a biological system [17]. 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 sense, 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 gain a more detailed integrative view of the dynamics of cellular processes in a cell or an organism rich in biological and spatial-temporal information [18]. 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.* [13] point out because it helps researchers to interpret heterogeneous data more quickly. 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 [22].

In this paper, we report the preliminary results we obtained 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 and Machine Learning algorithms could fail to good abstract interdependencies between the objects [7]. This fact often does not allow to generates a clear visual representation of the information.

Our idea is to show how a dynamic graph generation aided by a human can help abstract functional relationships between proteins to generate a clear data visualization where a standard clustering algorithm fails. For this contribution, we focused on two diseases: *Alzheimer* and *Parkinson*.

Alzheimer's disease (AD) is a form of degenerative dementia that occurs after 65 years. In this pathology, there is a deposition of an Aβ peptide B with the formation of senile plaques and the intracellular aggregation of *tau* protein [4]. Parkinson's disease (PD) is the second most com-

DOI reference number: 10.18293/DMSVIVA2021-013

mon neurodegenerative disorder in the senile age in which neuronal loss is found in the substance nigra and formation of α-synuclein aggregates that are neuropathological [14]. These pathologies show similar neurodegeneration mechanisms supported by scientific evidence with genetic, biochemical, and molecular studies. Pathological pathways involving α-synuclein and *tau* proteins, oxidative stress, mitochondrial dysfunction, iron pathway, and *locus coeruleus* are among these findings [21]. They were chosen as an example for our search workflow because of this overlapping between their pathological mechanisms. This aspect introduces intra-class, and extra-class overlaps able to fool standard clustering methods.

The paper is structured as follows. In Section 2 we describe the most important related works in the examined field. In Section 3 we explain the conducted experiments and in Section 4 we discuss respectively the dataset, methodology, and performance measures which we have used in our research. Finally, we expose the visual results in Section 5 and the conclusions with future work 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 that supports the standardization of biological information about attributes of genes and gene products through the use of ontology. It is structured as an acyclic oriented graph where each GO-term is identified by a word or strings and a unique alphanumeric code [6].

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 [2].

GOrilla[1] 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 [5].

Blast2GO (B2G)[2] 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 [3].

NaviGO[3], 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 [20].

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 [24].

3 Experimental setup

We explored two ways to calculate semantic similarity. We calculated the similarity for all three ontology gene domains, both for Alzheimer's and Parkinson's proteins, separately. For this first experiment, 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 about BPs, CCs, and MFs domains for AD and PD with the K-means algorithm, trying with n=3 and n=5 clusters.

4 Methods

In this work, we have used the R environment[4], a free software environment for statistical computing and graphics, and SigmaJS, a JavaScript library dedicated to graph drawing[5]. We used the standard SigmaJS renderer to show the graph view.

4.1 Datasets

Protein datasets for AD and PD were downloaded from UNIPROT [16]. Data cleaning has been carried out, removing all duplicates. Furthermore, for each UNIPROT protein

[1] Gorilla: http://cbl-gorilla.cs.technion.ac.il

[2] Blast2GO: https://www.biobam.com

[3] NaviGO: https://kiharalab.org/web/navigo/views/goset.php

[4] R: https://www.r-project.org

[5] SigmaJS: https://sigmajs.org

ID, the reference gene has been obtained and linked to the STRING, removing all the proteins which were not mapped in this database. STRING database allows us to consider any protein-protein interactions (PPI) based on a score calculated on experimental evidences [15]. We have recovered a total of 216 genes for AD and 137 genes for PD.

4.2 Gene Ontology

The Gene Ontology is based on two types of relationships between objects: *instances* and *part of*. Three considered all the organisms share biological domains and that constitute 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 with the following acronyms: biological process (BP), cellular component (CC), and molecular function (MF). We have recovered from UNIPROT[6] all the GO terms belonging to these three fields both for Alzheimer's and Parkinson's diseases with `UniProt` package in R.

4.3 Distance Metrics

We used two types of metric to calculate pairwise semantic similarities: *Lin* and *Wang* similarities with the `GOSemSim` package in R [23].

4.3.1 Lin's measure

Lin measure is based on *information content* (IC). The negative log of a concept's probability is formally known as information content (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 [10].

In this case, this approach can measure the similarity of the knowledge content of the GO terms for each protein dataset referring to the two diseases. The frequency of two GO words involved and their closest common ancestor in a particular corpus of GO annotations are used in the estimation. The most basic definition two concepts share as an ancestor is suggested by the term *Least Common Subsammer* (LCS) [12]. So, we can consider the following Equation 1:

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

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 measured across all proteins in the pathological reference dataset, both for AD and PD.

4.3.2 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 [19].

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:

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

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 below:

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

4.4 K-means

K-means is one of the most common and widely used partitioning clustering algorithms because it divides a set of objects into K clusters based on their attributes [11]. 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. As the name implies, a centroid is a real or imaginary point that represents the cluster's center and is updated with each algorithm iteration.

The procedure is composed by four steps:

- *Step 1:* determine the value of K;

[6] UniProt: https://www.uniprot.org

- *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:

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

  where dist (.) represents the Euclidean distance;
- *Step 4:* recalculate the updated cluster centers by averaging the points associated with each cluster:

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \tag{5}$$

  where Si 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. Due to the goal of this research, we limited the max number of clusters to five. No PCA techniques were used. Figures 2 and 3 report how the GO objects are partitioned regarding the BP features for both diseases. The axis reports the distance between each item to its centroid. We used `cluster` and `factoextra` packages in R to perform clusterization. It is hard to read this kind of visualization due to arbitrary parameter values that the final user can assign to the number of the cluster.

4.5 Dynamic Distance-Graph

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 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 Fig. 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,

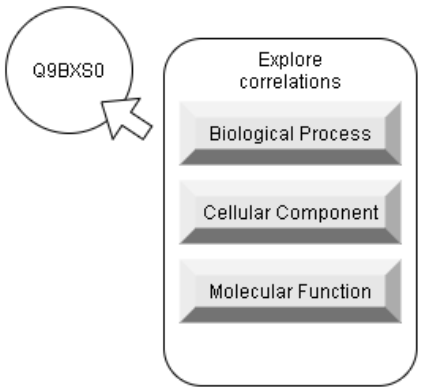


Figure 1: The contextual menu is available for each node.

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. Also, the distance value is used to separate nodes into spaces.

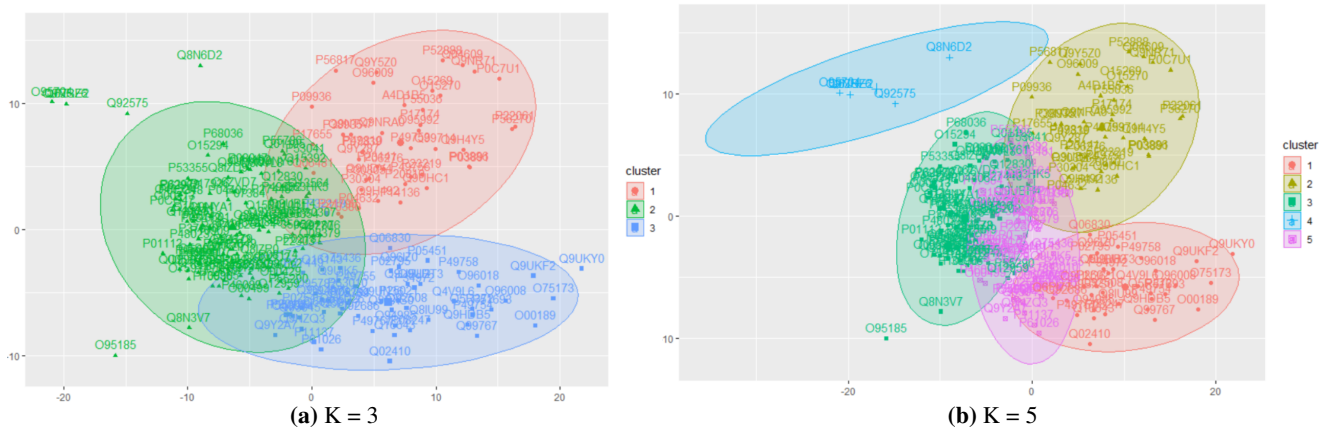
The ForceAtlas2 algorithm is used to avoid overlapping between near nodes. In particular, we used ForceAtlas2 embedded into SimgaJS [9]. ForceAtlas2 is a layout algorithm for force-directed graphs. This algorithm allows us to position each node depending on the other nodes using the distances between 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. Given two nodes, A and B and their own distance d:

$$SD = log_e(d) \tag{6}$$

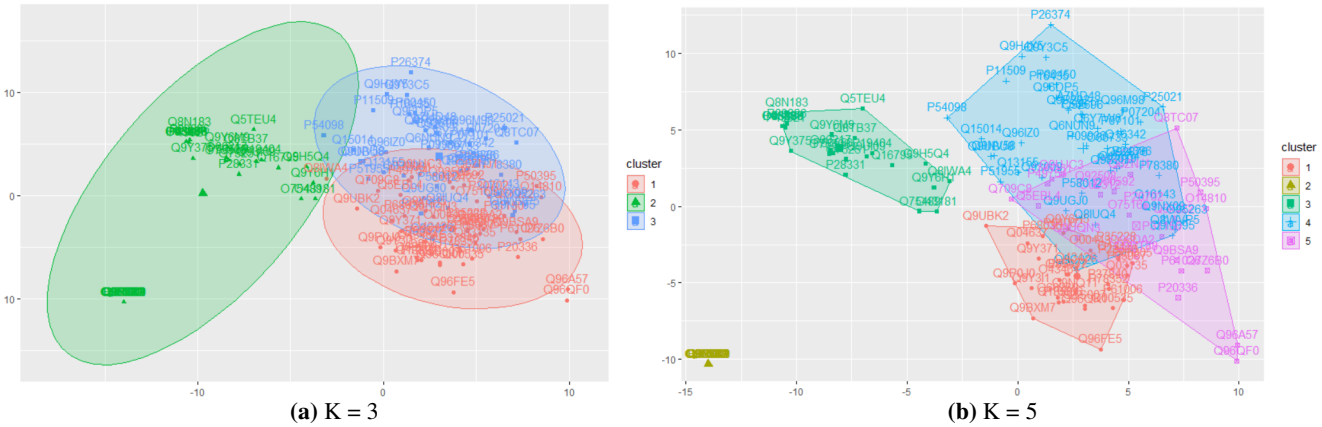
where *d* is the distance and the log_e is the natural logarithm with base the number of Nepero.

Note that SD is used only for graphical purposes in the rendering routines. Figure 5 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 5), the threshold th_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.



(a) K = 3 **(b)** K = 5

Figure 2: K-means for BP for Alzheimer's disease with Lin's measure.



(a) K = 3 **(b)** K = 5

Figure 3: K-means for BP for Parkinson's disease with Lin's measure.

5 Results

5.1 K-means visualization

We represented the images related to Alzheimer's and Parkinson's diseases and calculated with Lin's measure for convenience. We have found that clustering with the K-means algorithm produces visually misleading and uninformative overlaps. In Figures 2 and 3 is shown what happens when the data of Alzheimer ans Parkinson diseases for only BP component is clustered with n=3 and n=5.

5.2 DCDG visualization

For our test, we considered the G9BXS0 protein from the similarity matrices obtained and we identified the proteins of his neighbor to build our view of node expansion. Before testing DCDG view, we carried out a simple statistic of the common GO terms, even in this case for the only BP component, between this *root* protein and its neighbors. We represented them with a Venn diagram [8] (see Fig. 4), on the basis of GO Lin's similarity matrix.

This view allows us to evaluate which elements are common among the different sets of the terms GO for all the selected proteins. It is clear that a simple statistics of the terms does not make useful information beyond the simple observation that there are terms, even if minimal, common to all five sets of the Terms Go of each protein. Instead, introduce similarity based on the *information content* of the

Figure 4: Venn Diagram for G9BXS0.

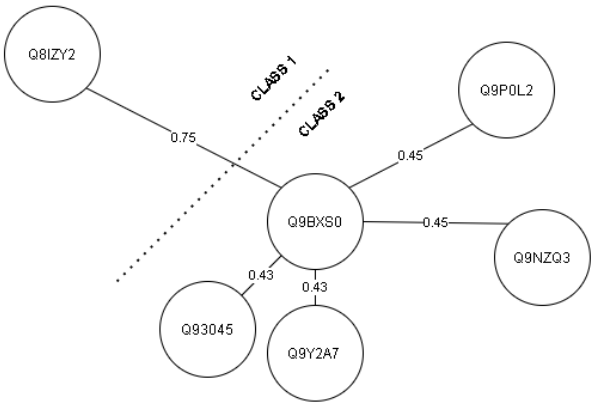


Figure 5: The result of Q9BX80 expansion by BP dataset.

GO terms is useful for expanding knowledge regarding biological aspects that would be omitted by a simple statistical analysis.

Figure 5 shows the BP expansion related to node G9BXS0, a protein produced by *COL25A1* gene for *Homo Sapiens* organism with the DCDG view. This protein inhibits the fibrillization of β-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 6 highlights the successive expansion of Q8IZY2 and Q9P0L2 proteins. Due to distances, a new class was identified by the system (**CLASS 3**). The visualization clearly states, from a point of view of biological meaning, that the added 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.

6 Conclusion

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

We used two similarity methods on the three GO vocabularies (*Biological Process, Cellular Component, and Molecular Function*) for two neurodegenerative diseases, Alzheimer's and Parkinson's diseases: Lin's and Wang's Methods. 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 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 for exploring protein data based on the proposed assumptions in this research study, guaranteeing user-target customization of the tools available.

References

[1] M. Arif. Similarity-dissimilarity plot for visualization of high dimensional data in biomedical pattern classification. *Journal of Medical Systems*, 36(3):1173–1181, 2012.

[2] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'donovan, and R. Apweiler. QuickGO: A web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

[3] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

[4] C. Duyckaerts, B. Delatour, and M.-C. Potier. Classification and basic pathology of Alzheimer disease. *Acta Neuropathologica*, 118(1):5–36, 2009.

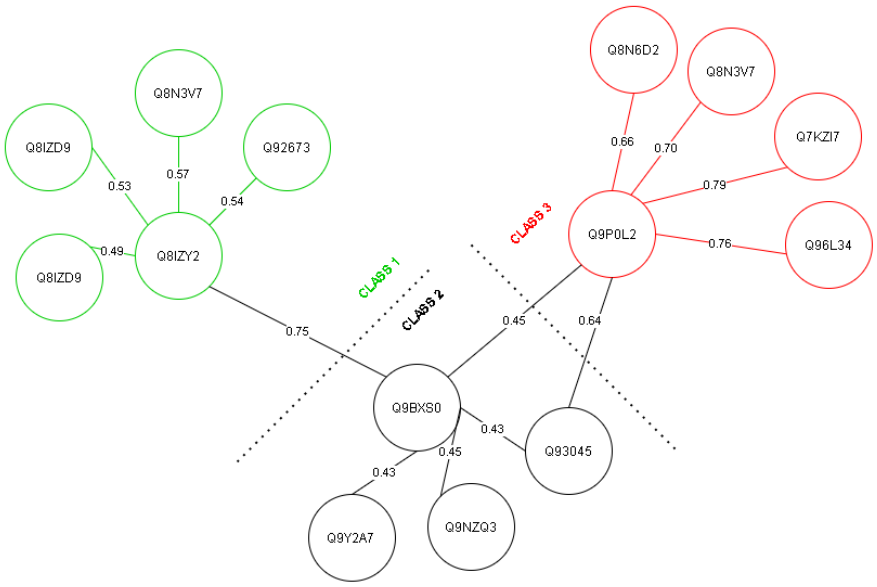


Figure 6: The result of Q8IZY2 and Q9P0L2 expansion by BP dataset.

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

[6] Gene Ontology Consortium. The gene ontology project. *Nucleic Acids Research*, 36(suppl_1):D440–D444, 2008.

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

[8] D. W. Henderson. Venn diagrams for more than four classes. *The American Mathematical Monthly*, 70(4):424–426, 1963.

[9] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6):e98679, 2014.

[10] D. Lin. Extracting collocations from text corpora. In *Proceedings of the First Workshop on Computational Terminology*, pages 57–63, 1998.

[11] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[12] B. T. McInnes and T. Pedersen. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, 46(6):1116–1124, 2013.

[13] S. I. O'Donoghue, A.-C. Gavin, N. Gehlenborg, D. S. Goodsell, J.-K. Hériché, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, et al. Visualizing biological data—now and in the future. *Nature Methods*, 7(3):S2–S4, 2010.

[14] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang. Parkinson disease. *Nature Reviews Disease Primers*, 3(1):1–21, 2017.

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

[16] UniProt Consortium. UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.

[17] M. Vailati-Riboni, V. Palombo, and J. J. Loor. What are omics sciences? In *Periparturient Diseases of Dairy Cows*, pages 1–7. Springer, 2017.

[18] T. D. Veenstra. Omics in systems biology: Current progress and future outlook. *Proteomics*, 21(3-4):2000235, 2021.

[19] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, 2007.

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

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

[22] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Briefings in Bioinformatics*, 19(6):1370–1381, 2018.

[23] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

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