EPiC Series in Computing

Volume 92, 2023, Pages 58–71

Proceedings of International Conference on Bioinformatics and Computational Biology (BICOB-2023)

Utilizing Functional Annotation of Disease Genes for Disease Clustering

Hisham Al-Mubaid[1] and Tamer Aldwairi[2]
[1] University of Houston – Clear Lake, Houston, TX, USA.
[2] Temple University, Philadelphia, PA, USA.
Hisham@UHCL.edu, aldwairi@temple.edu

Abstract

We investigate the task of disease clustering with the functional annotations of disease genes from the Gene Ontology using the biological process aspect. As an unsupervised machine learning step, the clustering task places communities of similar diseases together based on their closeness to one another using functional annotations of their associated genes. The research work and studies for the similarity, relationship, or clustering of human diseases using the functional information associated with the disease genes are limited. This work builds on and benefits from the advances in gene disease association studies; also from the advances in the functional annotations of human disease genes from the Gene Ontology. We validated the experimental results by comparing the intra-cluster and inter-cluster disease similarity with their semantic similarity in the is-a hierarchy in both MeSH and DO disease ontology. The experimental results are highly encouraging and show that we can rely on the functional profiles using the biological process annotations of disease genes for the study of disease clustering and similarity.

1. Introduction

Disease clustering, like disease similarity, is one of the important bioinformatics tasks for understanding various disease mechanisms at the molecular and functional levels [1, 2, 4, 5]. Disease clustering can be used to analyze the relationship among various diseases, specifically human diseases, by placing the diseases that are most similar together in the same group or cluster [2, 4]. Essentially, clustering is like classification a machine learning task that works on a fairly large set of data points

provided by the application or problem at hand. While classification is supervised, the clustering task is an unsupervised learning task that does not require prior labels for the data [2, 4, 5].

In this paper, we use the functional annotations of disease genes from the *Gene Ontology* (GO) for disease clustering. Specifically, we use the gene ontology functional annotation with *biological process* (bp) terms assigned to the various genes associated with human diseases for disease clustering. Most previous work on disease similarity and disease clustering relies on various kinds of disease information and attributes like disease symptoms, shared chemicals, disease genes, gene expression profiles, gene pathways, protein-protein interactions, and more [2, 4, 5, 6, 7]. In this work, we use both MeSH ontology and Disease Ontology (DO) for analyzing the semantic relationships among the diseases [8, 9, 11]; however, we rely more on the MeSH ontology [8] {*note: besides Bioportal we also used MeSH disease data from file CTD_diseases which includes parent ID for every disease* [21]}. Further, we applied the semantic relationship between diseases from MeSH and DO to examine and validate the outcomes of the proposed disease clustering method. The evaluation results are encouraging and prove that the functional profiles of diseases from the *biological process* (bp) aspect of the GO are a good attributes for disease similarity and clustering. The main contribution of this work is the investigation of the utility and benefit of using the gene ontology functional annotation of human genes in the study of disease relationships represented by disease clustering. This kind of work has never been investigated extensively to the best of our knowledge. We examined a fairly good number of evaluation settings for disease clustering by utilizing only the disease functional annotations from the bp aspect of the gene ontology.

2. Background and Related Work

The important and most commonly used clustering algorithms include: *Agglomerative Hierarchical clustering*, *k-means clustering* and *Dbscan* [3, 12, 13]. While k-means clustering is categorized in the centroid-based methods, Dbscan is categorized in the density-based clustering methods, and the Agglomerative Hierarchical clustering is categorized in the hierarchical clustering methods [5].

In [10], Godwin and Ugwoke (2018) applied clustering in the healthcare field to identify groups of patients with diabetes who have similar profiles(e.g., age and gender) as well as common clinical histories [10]. Bello et. al. (2018) discussed the importance of the adoption of Disease Ontology which will help in unifying disease annotations across different species through the collaborative effort of aligning disease terms across different projects [22]. They emphasize that the collaboration between the Mouse Genome Database, the Rat Genome Database, and the Disease Ontology project demonstrates the usability of Disease Ontology across different model organisms within the database community [22].

In [5], Karim et. al (2021) presented an extensive study of clustering and cluster analysis that is based on representation learning for helping bioinformatics research. They reviewed most of the deep-learning-based clustering approaches. Their evaluation was conducted with three bioinformatics tasks: bioimaging, cancer genomics and biomedical text mining. [5].

The Dbscan clustering is one of the most reliable clustering techniques for large data sets with different sizes and arbitrary shapes. The main focus of the algorithm is on finding the densest areas and recursively extending them to find dense arbitrarily shaped clusters [5, 25]. Both Anand and Kumar (2018), as well as Karim et. al. (2021), present fairly comprehensive studies and surveys about the various clustering algorithms and their applications in various fields [5, 25].

3. Methods and Techniques

As an unsupervised machine learning task, the clustering task is the process of gathering similar data points into a predefined number of bins and each bin is called a *cluster*. A clustering algorithm, e.g., *Dbscan* [12, 13] or *Hierarchical clustering*, tries to learn some hidden patterns that can be used to group similar items together in a meaningful way. In this work, we utilize the functional annotations of disease genes for disease clustering. The Gene Ontology is composed of three aspects: *biological process* bp, *molecular function* mf, and *cellular component* cc. The bp aspect is a taxonomy of all bp functional terms that can be assigned for various gene products in various organisms [16, 17, 23]. Each human disease gene is annotated with one or more functional terms from the bp aspect. The database of all *gene-bp* functional assignments is the *Gene Ontology Annotation* (GOA) [18] which includes all GO-gene annotations and considered the most comprehensive dataset for gene ontology annotations (including *bp* annotations) of human genes [16 – 18]. Each disease is represented as a vector of *bp* terms and the clustering method assigns diseases to clusters based on their bp profiles from their associated genes. That is, diseases that are annotated with similar sets of *bp* functions will be assigned (or placed) in the same cluster. Two diseases d_x and d_y do not need to be associated with the same disease genes in order to have similar sets of *bp* terms (note: we use the words '*bp term*' and '*bp function*' interchangeably for the same meaning to indicate one node or term in the bp taxonomy}. For example, for some disease d_x let the set $g(d_x)$ be the set of all genes associated with the disease d_x: $g(d_x) = \{g_1, g_2, g_3\}$ and similarly $g(d_y) = \{g_4, g_5\}$; therefore, here the two diseases d_x and d_y have completely different genes associated with them. Now suppose the set $p(d_x)$ be the set of bp terms associated with disease d_x; and also let: $p(d_x) = \{t_1, t_2, t_3, t_4, t_5\}$ and $p(d_y) = \{t_2, t_4, t_5, t_6\}$. As we can see, the two diseases $d_x$and d_y have very similar sets of bp terms (*the three bp terms* t_2, t_4, t_5 *are in common*) even though their associated genes are completely different .

Once all the disease pb term assignments are constructed into disease vectors we use the $Dbscan$ clustering algorithm [12 – 14] to cluster the diseases based on their biological process functional annotations. The *Density-Based Spatial Clustering of Applications with Noise* (Dbscan) is the clustering algorithm we employ in this work. Dbscan is one of the most commonly used and well-known clustering algorithms for similar tasks like these. Based on the distance measure between the data points, Dbscan puts together (in the same cluster) all data points that are close to each other. We used disease information and disease data from the following sources: - OMIM for disease information in OMIM and *morbidmap* [19]; *DO* (The Disease Ontology) [11, 20] for the hierarchical relationship between diseases; and for *MeSH* disease info we used the MeSH db from BioPortal [8, 9]; we used the GOA_human for the gene ontology bp functional annotations of human disease genes [18]. Finally, we used the *CTD* database [21] for disease data, gene-disease associations, and hierarchical relationships (parent-child) between *MeSH* diseases [21]. For example, from the file CTD_diseases: we extracted parent diseases for each disease with MeSH disease ID which also has MeSH Parent IDs, as follows:

Disease Name	Disease ID	Parent IDs
Alzheimer Disease	MESH:D000544	MESH:D003704 , MESH:D024801
Explained: This records shows that the Alzheimer disease (disease Id in MESH: D000455) has two parents: (1) Dementia MESH:D003704 and (2) Tauopathies MESH:D024801		

And these two diseases (*Dementia and Tauopathies*) has the following parent nodes:

Disease Name	Disease ID	Parent IDs
Dementia	MESH:D003704	MESH:D001927 , MESH:D019965
Tauopathies	MESH:D024801	MESH:D019636

For each disease d_i we obtained from *morbidmap* [19] all the genes associated with d_i (*disease-genes*); also this information can be verified from the *CTD* database [21]. Next, we used the *GOA_human* data [18] (*which includes 564,813 GO annotations for all human genes*) from the *Gene Ontology* to extract all *biological process* (bp) function terms associated with each gene [18]. Then, we constructed the disease vector for each disease from the *bp* function terms of its genes. Let v_i be the vector for disease d_i such that v_{ij} is the *jth* component of the vector which represents bp term p_j such that v_{ij}=1 if disease d_i is annotated with p_j and 0 otherwise:

$$v_{ij} = \begin{cases} 1 & if\ d_i\ is\ annotated\ with\ term\ pj \\ 0 & otherwise \end{cases}$$

For example: let $v_t = \{1, 0, 1, 1, 0, 1\}$ be the disease vector for some disease d_t and this disease is annotated with the bp terms p_1, p_3, p_4 and p_6. If a disease is annotated with the *bp* term p_i then it will be also annotated with all the ancestors of p_i in the biological *process taxonomy* to the root term (*the root term for the bp taxonomy is biological_process and its GO Id is GO:0008150*).

We constructed a superset of *bp* terms to be used as feature space for the disease bp vectors as follows:

- s1: From OMIM morbidmap we selected 1000 diseases from disease marked with (3) to indicate *molecular basis known*. Notice that from the total of *8,218* diseases in the OMIM *morbidmap*, there are *7,302* diseases marked as ' *# 3 - The molecular basis for the disorder is known; a mutation has been found in the gene*.'; (all the *8,218* diseases are listed with their associated genes). Furthermore, we found in *morbidmap* a total of *5,962* unique (*human*) genes.
- s2: Then we obtained all genes associated with each disease in step s1; we call this set *dgss* (disease genes superset).
- s3: From the *GOA_human* we obtained all bp terms annotated for all disease-genes in the set *dgss* (from step s2 above); we call this set *pbss* (bp superset).
- s4: We removed from *bpss* every bp term that is annotating ≥90% or ≤ 10% of the *dgss* set.

The resulting pb terms superset *bpss* is then applied as the feature space for disease feature vectors v_i.

4. Results and Discussion

We conducted several evaluations to test our disease clustering approach with different numbers of diseases and various number of clusters in multiple experimental settings. Table 1 contains 100 diseases selected for one of the experiments from *MeSH* and also the *morbidmap* in OMIM[19].

Evaluation 1: The first three experiments *e1* to e3 were conducted with 20, 30 and 50 diseases respectively. Then each experiment is repeated two times with two settings by changing k (k : the number of clusters) as *k=2* and *k=3*; as follows:

Experiment e1: we randomly selected 20 diseases from *morbidmap* data and with two clustering's k=2, k=3 and this resulted in experiments e1a and e1b; as shown:

Experiment	# of diseases	k (# of clusters)
e1a	20	2
e1b		3

Experiment e2: Similarly, experiment e2 was done with 30 diseases and repeated two times by changing number of clusters as follows:

Experiment	# of diseases	k (# of clusters)
e2a	30	2
e2b		3

Experiment e3: like the previous two experiments, we selected 50 diseases for this experiment with two different number of clusters as follows:

Experiment	# of diseases	k (# of clusters)
e3a	50	2
e3b		3

For these experiments (*e1 – e3*) we evaluated the clustering quality by comparing the path length between two disease nodes in the *MeSH* and *DO* disease ontology [11, 20] using edge count. {*note: we used the BioPortal for shortest path length and edge count for both MeSH and DO; we also utilized the Parent ID in the CTD database [21] file CTD_diseases which includes every is-a relationship between a MeSH disease and its parent disease*}. For example, in Figure 1, the path length between the two disease nodes {*heart disease*} and {*double outlet right ventricle*} is 2 (*by counting the edges between these two nodes*). If there are more than one path between two nodes in the ontology, we select the shortest path. For example, the two disease nodes {*disease*} and {*double outlet right ventricle*}in Figure 1 have four different paths between them with the shortest path length is being 3 (*the other three paths are of length 4 and 5 using edge counting*). In another example, in Figure 2 (a portion of the MeSH ontology); the path length (PL) between *Alzheimer Disease* and *Nervous System Diseases* is 3 while the PL between *Alzheimer Disease* and *Dementia* is 1 (which indicates that *Alzheimer Disease* and *Dementia* are highly similar). An ontology, such as DO disease ontology or MESH, is a manifestation of the semantic relationship (*is_a* relationship) between the nodes of the ontology. For example, as shown in Figure 8, in the DO ontology, the disease Cataract 20 multiple types is_a Cataract.

In general, the shortest path between any two term nodes in a given ontology (*a Directed-Acyclic Graph DAG*) has been used as a measure of similarity (or relatedness) of the two terms [15]. Given that the path length can be a measure of relatedness between diseases, we computed the shortest path length between every disease pairs within one cluster (*intra-cluster*) and computed the average of all pairs in the cluster. Then we computed the shortest path length between every disease pair with two diseases from two different clusters (*inter-cluster distance*). The results are shown in Table 2 for the first three experiments *e1 – e3* and illustrated in Figure 3.

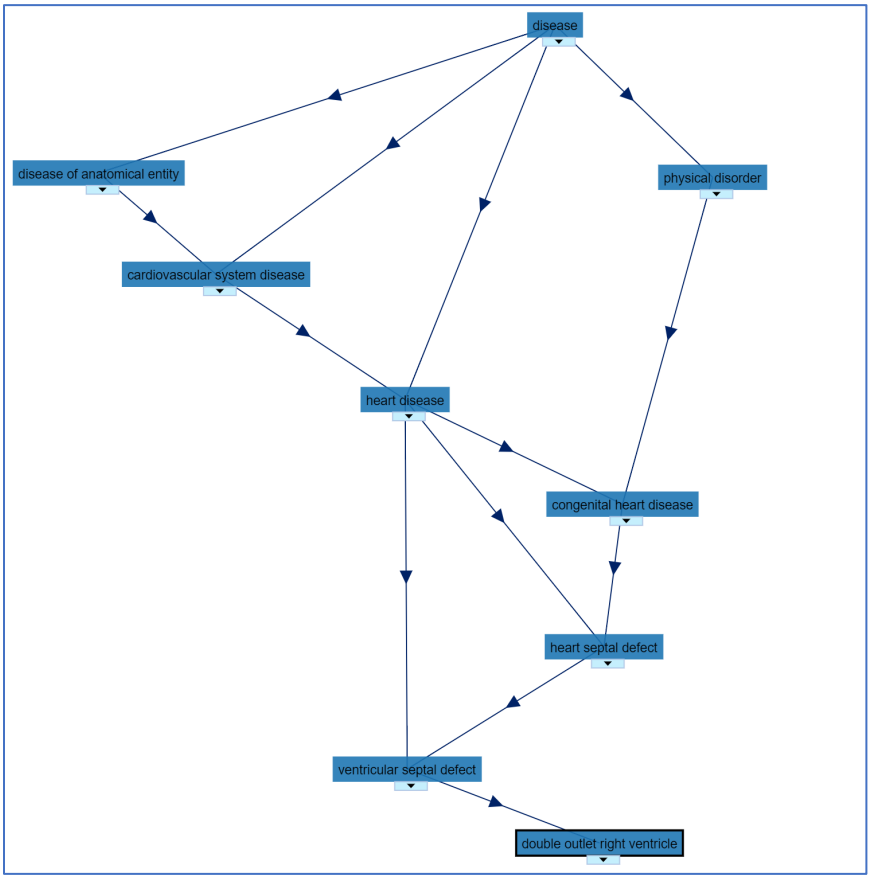

Figure 1: portion of the DO for disease name *double outlet right ventricle* DOID:6406 {OMIM:217095, MESH:D004310 }

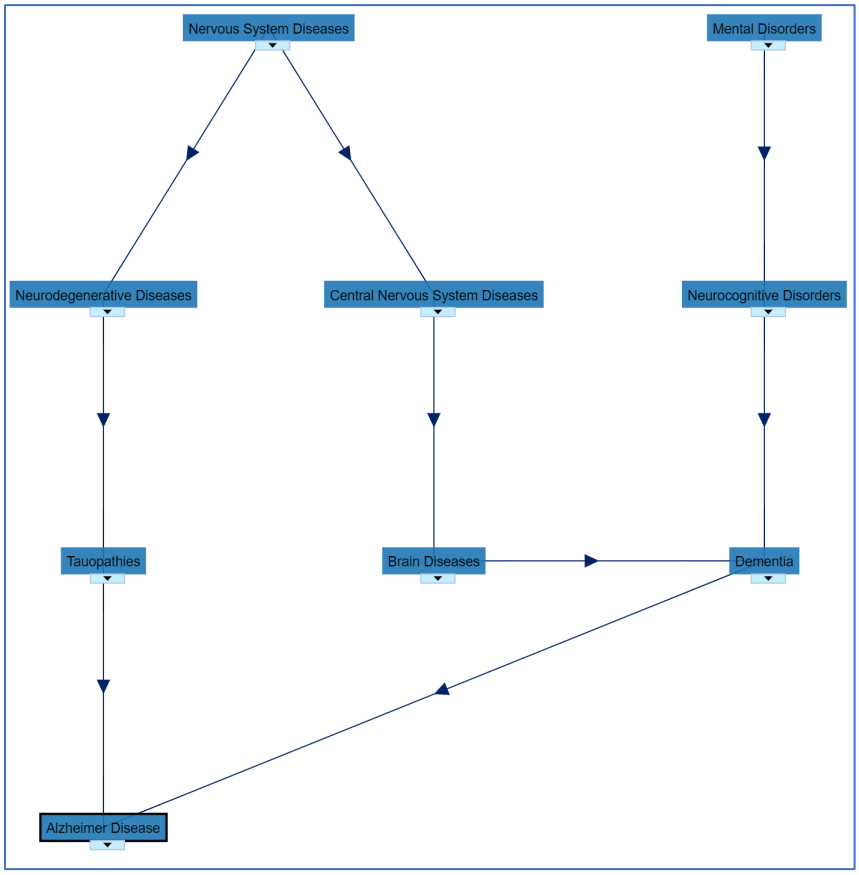

Figure 2: portion of the MESH ontology showing the nodes from Alzheimer Disease to the root.

Experiment	Intra-cluster distance {path length: # of edges in MESH taxonomy}	Inter-cluster distance {path length: # of edges in MESH taxonomy}
e1	4.85	6.35
e2	3.41	4.82
e3	3.65	5.05

Table 2: The mean value for the shortest path length between disease nodes within a cluster and between clusters.

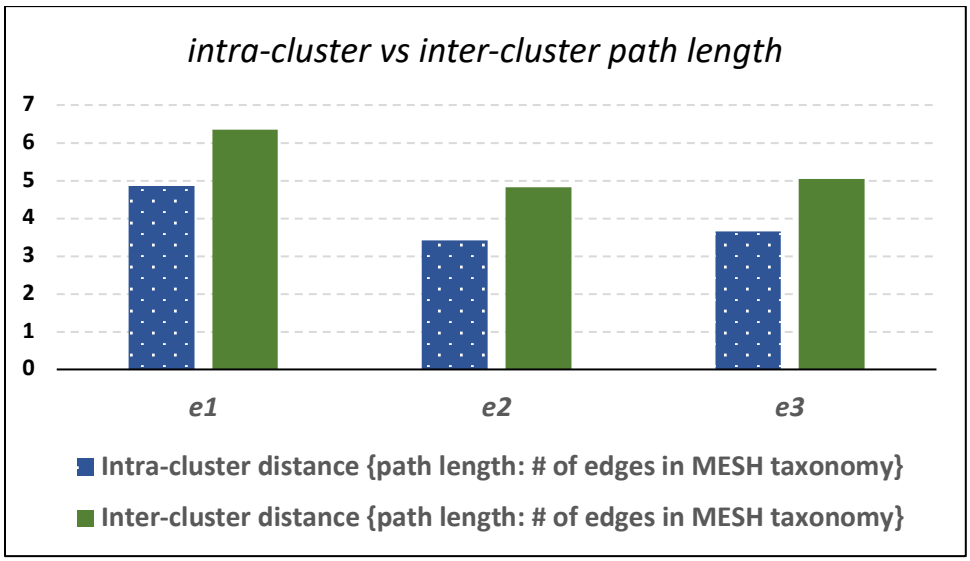

Figure 3: This figure shows the difference between average path length between disease pairs within one cluster (intra-cluster) versus disease pairs from two clusters (inter-cluster).

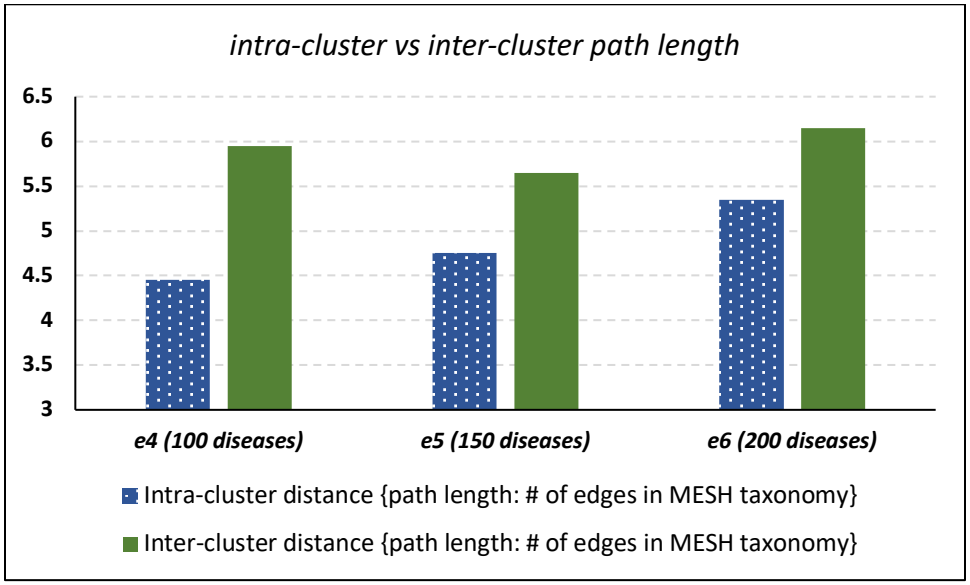

Figure 4: This figure shows the average path length between disease pairs within one cluster (intra-cluster) versus disease pairs from two clusters (inter-cluster) for the second evaluation with 100 – 200 diseases and two clusters (k=2).

Hypomagnesemia 5, renal, with ocular involvement, 248190	{Psoriasis susceptibility 5}	Reducing body myopathy, X-linked 1b, with late childhood or adult onset, 300718
Cockayne syndrome, type B, 133540	Cardiomyopathy, hypertrophic, 17, 613873	Hemorrhagic diathesis due to antithrombin Pittsburgh, 613490
3-methylglutaconic aciduria, type V, 610198	?Dyskeratosis congenita, autosomal recessive 7, 616553	Intellectual developmental disorder, autosomal dominant 51, 617788
Hair, curly	Thyroid dyshormonogenesis 3, 274700	Spinocerebellar ataxia 25, 608703
Pseudohypoaldosteronism, type IB3, autosomal recessive, 620126	Adams-Oliver syndrome 5, 616028	Deafness, autosomal recessive 12, 601386
{Hypercalciuria, absorptive, susceptibility to}, 143870	Langer mesomelic dysplasia, 249700	Emery-Dreifuss muscular dystrophy 1, X-linked, 310300
Norum disease, 245900	Myopathy, distal, with anterior tibial onset, 606768	{Synovitis, chronic, susceptibility to}
Ehlers-Danlos syndrome, arthrochalasia type, 1, 130060	?COACH syndrome 3, 619113	{Macular degeneration, age-related, 2}, 153800
{Bone mineral density QTL 12, osteoporosis}, 612560	Muscular dystrophy-dystroglycanopathy (congenital with impaired intellectual development), type B, 6, 608840	Keratoderma, palmoplantar, punctate type IB
Kilquist syndrome, 619080	?Megaloblastic anemia, folate-responsive, 601775	Bartter syndrome, type 4b, digenic, 613090
{Diabetes, type 2}, 125853	?Neurodegeneration with brain iron accumulation 7, 617916	Alazami-Yuan syndrome, 617126
Gallbladder disease 3	Ciliary dyskinesia, primary, 20, 615067	Cardiomyopathy, hypertrophic 6, 600858
Spastic paraplegia 14, autosomal recessive	Aminoacylase 1 deficiency, 609924	Seizures, scoliosis, and macrocephaly syndrome, 616682
Cardiomyopathy, dilated, 1MM, 615396	Bardet-Biedl syndrome 8, 615985	{Autism susceptibility 4}
?Acne inversa, familial, 3, 613737	Febrile seizures, familial, 5	Holocarboxylase synthetase deficiency, 253270
Autoimmune lymphoproliferative syndrome, type III, 615559	Congenital myopathy with excess of muscle spindles, 218040	Orofacial cleft 8, 618149
Dystonia 4, torsion, autosomal dominant, 128101	Cone dystrophy 4, 613093	Stuttering, familial persistent, 2
?SERKAL syndrome, 611812	Acromelic frontonasal dysostosis, 603671	Mitochondrial DNA depletion syndrome 4B (MNGIE type), 613662
Microphthalmia, isolated 6, 613517	Cervical cancer, somatic, 603956	{Prostate cancer, familial, susceptibility to}, 176807
Spinocerebellar ataxia 5, 600224	?Caudal duplication anomaly, 607864	van Buchem disease, type 2, 607636
Hypercalcemia, infantile, 1, 143880	Noonan syndrome 10, 616564	{Efavirenz central nervous system toxicity, susceptibility to}, 614546
Spinal muscular atrophy, infantile, James type, 619042	?Lipodystrophy, congenital generalized, type 3, 612526	Cone dystrophy-3, 602093
Tooth agenesis, selective, X-linked 1, 313500	[Mean platelet volume QTL3]	[Skin/hair/eye pigmentation 1, blue/nonblue eyes], 227220
[Blood group, Kell], 110900	{Schizophrenia}, 181500	Thyroid carcinoma, papillary, with papillary renal neoplasia
Keratolytic winter erythema, 148370	Spinal muscular atrophy-2, 253550	Spastic paraplegia 10, autosomal dominant, 604187
?Bleeding disorder, platelet-type, 22, 618462	[Uric acid concentration, serum, QTL5]	Pseudoxanthoma elasticum, 264800
Macular dystrophy, vitelliform, 3, 608161	?Hypertrichosis, congenital generalized, with gingival hyperplasia, 135400	{Dermatitis, atopic, susceptibility to, 3}
{Asthma, susceptibility to, 2}, 608584	Nystagmus 7, congenital, autosomal dominant	Verheij syndrome, 615583
{Hemolytic uremic syndrome, atypical, susceptibility to, 2}, 612922	{Low renin hypertension, susceptibility to}	Leukemia, acute myeloid, 601626
Biliary cirrhosis, primary, 5	Gustavson syndrome	Pulmonary venoocclusive disease 2, 234810
Heart-hand syndrome, Slovenian type, 610140	Lymphatic malformation 2	Cystinuria, 220100
Olmsted syndrome 2, 619208	Developmental and epileptic encephalopathy 16, 615338	Cleidocranial dysplasia 2, 620099
Gastric cancer, somatic, 613659	?Moebius syndrome	
Dentici-Novelli neurodevelopmental syndrome, 619877	Brooke-Spiegler syndrome, 605041	

Table 1: Sample of 100 diseases selected from the morbid map file.

Evaluation 2: Then, after the first three experiments (*e1 – e3*) in *Evaluation 1*, we increased the number of diseases. Therefore, we conducted another group of experiments with number of diseases 100, 150 and 200 with experiments *e4*, *e5* and *e6* respectively and we kept k=2 clusters; the results are shown in Figure 4.

Evaluation 3: in this evaluation, we selected two groups g_1 and g_2 of diseases based on the shortest path length between them in the DO ontology. Each group consists of 20 diseases and *n=40* (n: total number of diseases). The shortest path length between diseases in each group ranges between 1 and 4:

g_1 : 20 diseases; shortest path length between them: 1 – 4; each disease d_i is labeled with $l_i = 1$
g_2 : 20 diseases; shortest path length between them: 1 – 4; each disease d_j is labeled with $l_j = 2$
Moreover, the shortest path length between any disease pair from the two groups ranges between 6 – 8; as shown in Figure 5.

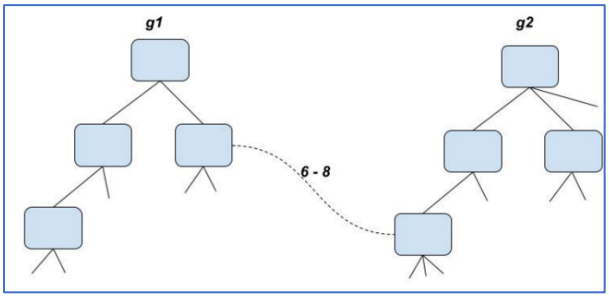

Figure 5: The shortest path length between any two nodes within g_1 (or within g_2) ranges between 1 – 4 while the shortest path length between any two disease from g_1 and g_2 ranges between 6 – 8.

Define c_i to be the cluster for disease d_i produced by the proposed clustering method. Then, for a given disease d_q if $c_q = l_q$ then the clustering method placed d_q in its correct group. With this, we can use the clustering accuracy to examine and validate the proposed clustering method as follows:.

$Accuracy = \sum_{i=1}^{n} clustering(d_i)$, where $clustering\ (d_i) = \begin{cases} 1 & if\ c_i = l_i \\ 0 & otherwise \end{cases}$

This setting is repeated ten times with each repetition conducted with different diseases in the two groups. This evaluation therefore is done with 400 diseases in ten rounds. The accuracy results of the ten rounds are sorted ascending accuracy and are illustrated in Figure 6. As shown in Figure 6, the accuracy ranges from 0.70 to 0.95 with average accuracy is 0.825 indicating that the clustering method was able to place (on average) 33 out of the 40 diseases in their correct groups.

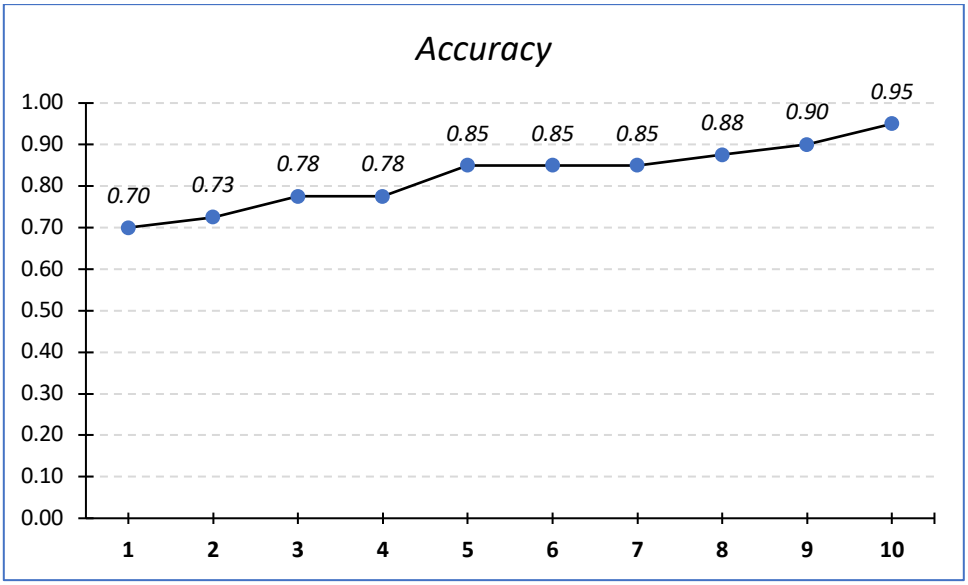

Figure 6: The accuracy of clustering assignment compared with the truth label l_i for each disease d_i.

Evaluation 4: In the last evaluation, we conducted experiments with 20 diseases each and k=*2*; this is repeated 20 times with a total of 400 diseases in 20 disease clustering experiments. In this evaluation,

we selected diseases that do not share any genes (no shared genes in each 20 diseases experiment). We wanted to verify the validity of the clustering based solely on the bp functional profiles and independently from the genes associated with diseases. The results are shown in Figure 7. These results demonstrate clearly that there are noticeable differences in the average path length of intra-cluster versus inter-cluster disease pairs in all 20 experiments.

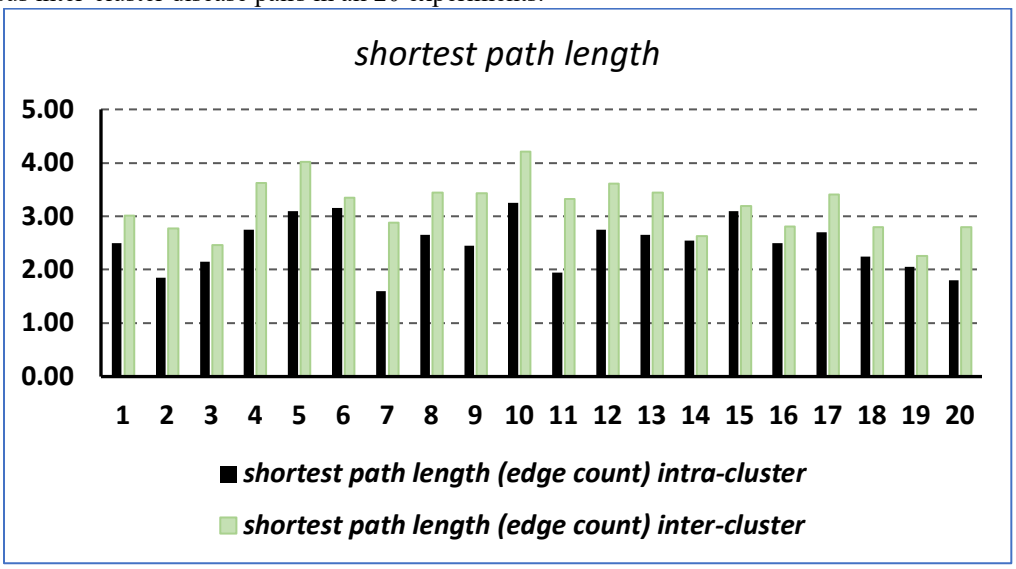

Figure 7: The average shortest path length of 20 clustering experiments each with 20 diseases and k=2. Each experiment includes 20 diseases that have no shared genes.

Discussion: We expect that if the clustering algorithm places two disease d_i and d_j in the same cluster then d_i and d_j are relatively more similar to each other than any two diseases taken from two clusters. Moreover, if the shortest path length $PL(d_i, d_j)$ between diseases d_i and d_j in a given disease ontology is less that the $PL(d_x, d_y)$ then the disease pair d_i, d_j is more semantically similar than the pair d_x, d_y. In the last evaluation, for example, disease pairs (d_i, d_j) in the same cluster (*i.e. the clustering algorithm places both d_i and d_j in the same cluster*) have average shortest path length between them =2.49; on the other hand, disease pairs (d_p, d_q) from two clusters (*i.e. d_p in one cluster and d_q in the other cluster)* on average have shortest path length = 3.17; as follows:

Cluster	No. of diseases (# of disease pairs)	Avg. shortest PL	
Cluster 1	12 (66 pairs)	2.44	2.49 (avg intra-cluster PL)
Cluster 2	8 (28 pairs)	2.62	
Inter-cluster	20 (96 pairs)	3.17	3.17 (avg inter-cluster PL)

This indicates that the clustering algorithm groups the semantically similar diseases together reliably according to the disease ontology.

5. Conclusions

We presented the method and results of disease clustering using the functional annotation of disease genes from the biological process aspect in the Gene Ontology. Disease clustering, like disease similarity, is an important task for understanding the various disease mechanisms and aspects at various molecular and functional levels. Also, disease similarity outcomes are valuable for disease relationship analysis, gene disease association, and drug repurposing studies. We conducted experiments with various number of diseases and clusters and with validation using both MeSH and the DO ontology. The results are highly encouraging. The proposed method groups similar diseases together based on their semantic similarity in the *is-a* hierarchy from *MeSH* or *DO* ontology by using the gene ontology *bp* annotations which is a completely different and independent information resource than both *MeSH* and *DO* ontology. The work in this paper overall shows that it is fairly reliable to use the gene ontology functional annotation of disease genes for disease clustering.

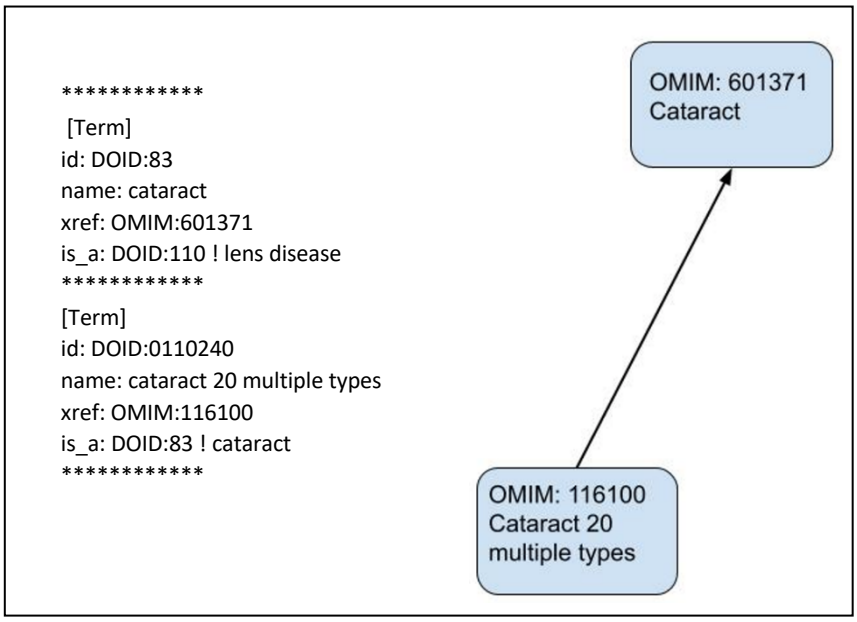

Figure 8: Example of two diseases with an is_a relationship (parent-child relationship) in the DO ontology.

References

[1] Halu, A., De Domenico, M., Arenas, A. et al. The multiplex network of human diseases. npj Syst Biol Appl 5, 15 (2019). https://doi.org/10.1038/s41540-019-0092-5

[2] Sanchez-Valle J, Tejero H, et. al. Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. Nature Communication 2020; 11 (1): 2854. doi: 10.1038/s41467-020-16540-x. PMID: 32504002;

[3] Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. Appl. Stat. 1979; 28:100.

[4] Rojano E, Córdoba-Caballero J, Jabato FM, et al. Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer. J Pers Med 2021;11(8):730.

[5] Md Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, Stefan Decker. Deep learning-based clustering approaches for bioinformatics. Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 393–415, https://doi.org/10.1093/bib/bbz170

[6] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning. New York City, NY, USA: ICMLR, 2016, 478–87.

[7] P. Dahal. Learning embedding space for clustering from deep representations. Proc. IEEE International Conference on Big Data (2018), pp. 3747-3755

[8] Leydesdorff L, Comins JA, Sorensen AA, Bornmann L, Hellsten I. Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level. Scientometrics. 2016;109(3):2077-2091. doi: 10.1007/s11192-016-2119-7. PMID: 27942085;

[9] BioPortal: https://bioportal.bioontology.org/ontologies/MESH/

[10] Ogbuabor Godwin and Ugwoke F. N.. 2018. Clustering algorithm for a healthcare dataset using silhouette score value. International Journal of Computer Science & Information Technology 10, 2 (2018), 27–37.

[11] Disease Ontology (DO): https://disease-ontology.org/

[12] M. Ester, H.P. Kriegel, et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise Int. Conf. Knowledge Discov. Data Min. (1996).

[13] Jiang Xie, Ruiying Wu, Haitao Wang, Haibing Chen, Xiaochun Xu, Yanyan Kong, Wu Zhang, Prediction of cardiovascular diseases using weight learning based on density information, Neurocomputing, Volume 452, 2021, Pages 566-575, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2020.10.114.

[14] Damodar Reddy Edla, Prasanta K. Jana. A Prototype-Based Modified DBSCAN for Gene Clustering. Procedia Technology, Volume 6, 2012, ISSN 2212-0173, https://doi.org/10.1016/j.protcy.2012.10.058.

[15] S.R. Maetschke, M. Simonsen, M.J. Davis, M.A. Ragan. Gene Ontology-driven inference of protein–protein interactions using inducers. Bioinformatics, Volume 28, Issue 1, 2012, Pages 69–75, https://doi.org/10.1093/bioinformatics/btr610

[16] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

[17] The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Jan 2021;49(D1):D325-D334.

[18] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot

with Gene Ontology. Nucleic Acids Res. 2004; 32 (Database issue):D262-6. doi: 10.1093/nar/gkh021. PMID: 14681408;

[19] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {Oct. 2022}. World Wide Web URL: https://omim.org/

[20] Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., … Greene, C. (2018). Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic acids research, 47(D1), D955–D962. doi:10.1093/nar/gky1032 PMID:30407550

[21] CTD: Comparative Toxicogenomics Database: https://ctdbase.org/

[22] Susan M. Bello, Mary Shimoyama, Elvira Mitraka, Stanley J. F. Laulederkind, Cynthia L. Smith, Janan T. Eppig, Lynn M. Schriml; Disease Ontology: improving and unifying disease annotations across species. Dis Model Mech 1 March 2018; 11 (3): dmm032839. doi: https://doi.org/10.1242/dmm.032839

[23] Ruths, T., Ruths, D. & Nakhleh, L. Gs2: an efficiently computable measure of go-based similarity of gene sets. Bioinformatics 25, 1178–1184 (2009).

[24] Liu M, Thomas PD.GO functional similarity clustering depends on similarity measure, clustering method, and annotation completeness. BMC Bioinformatics. 2019;20(1):155. doi: 10.1186/s12859-019-2752-2.

[25] Sanjay Kumar Anand and Suresh Kumar. 2022. Experimental Comparisons of Clustering Approaches for Data Representation. ACM Comput. Surv. 55, 3, Article 45 (March 2023), https://doi.org/10.1145/3490384