May-June 2020
ISSN: 0193-4120 Page No. 5124-5129

Evaluating Semantic Similarity and Centrality on Gene Annotation

[1]Aishwarya AV, [2]Anooja Ali, [3]Vishwanath R Hulipalled, [4]Akshita Srikanth, [5]Aishwarya Gajanana Naik,

[1]School of C&IT, REVA University, India, aishuav01@gmail.com
[2]School of C& IT, REVA University, India, anoojaali@reva.edu.in
[3]School of C&IT, REVA University, India, vishwanath.rh@reva.edu.in
[4]School of C&IT, REVA University, India, akshita291997@gmail.com
[5]School of C&IT, REVA University, India, gnaikaishwarya97@gmail.com

Article Info
Volume 83
Page Number: 5124-5129
Publication Issue:
May-June 2020

Abstract
Gene Ontology (GO) is a vocabulary available in bio informatics that indicates the functionality of proteins and genes. This dynamic vocabulary demonstrate the functionality at cellular component, biological process and molecular level. Different methods are there to evaluate this semantic similarity focusing on multiple approaches. In this paper we use jackknife methodology by considering five popular similarity measures. Protein Protein Interaction network (PPI) is created based on these similarity values, there by leading to the formation of clusters of identical or similar protein complexes. There are various methods available in literature to detect the essential proteins. These essential proteins are the hub nodes in the network. To form clusters of these networks, we apply various centrality measures to identify the most influential node. The clusters so formed help us in easy identification of the category of protein complex they belong to. Disease pathways are disintegrated and reasonably implanted in PPI network. So the research to discover the disease pathways over the set of predefined gene annotation can provide further advances in disease gene discovery.

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Keywords: *Centrality, Cluster, Gene Ontology, Protein-Protein Interaction (PPI) Network , Semantic Similarity.*

1. Introduction

Proteins play a major role in cellular functions. For understanding cellular processes, identification of essential proteins through *PPI* networks has great significance [1]. Identifying protein complexes help in understanding the building blocks of an organism and also the characteristics of proteins help in prediction of related disease or of target cells that might be associated with it [2].

Gene control the functions performed by protein. The process of identifying the coding regions in genes and their functionalities are called gene annotation. It helps in finding and attaching structural elements and its related function to each gene location. It has all the biological information to build any given living organism. With the help of gene annotation, it is possible to identify and predict the functions of protein complexes and thus perform further comparative analysis [3].

Essential proteins play an important role in maintaining cellular life. The common methods for essential protein detection include mutagenesis and gene knockout [4]. The native methods of identification of essential proteins were time consuming. Thus there has been a significant effort to discover complexes in *PPI*[5].

GO provides the function associated with genes and it is most widely used as a resource for gene annotations. *GO* is a bio informatics resource, can be

called as a biological vocabulary that display the functionalities of the main categories including Semantic Similarity is the similarity among structure and syntax of the sentences, it is considered that both the sentences convey a similar meaning. To give a numeric value (measuring value) to the *GO* term between the genes, semantic similarity of GO terms is performed [6].

The protein-protein interaction network is a graphical representation of proteins connected to each other through edges, it can be an undirected or directed graph [7]. The edges may contain weights representing the closeness between the protein nodes. Mapping similar genes/proteins to each other in a graph, the PPI networks can be formed. These networks usually depict a biological function.

In this paper we detect the functional similarity between the genes. The functionally similar genes are interconnected to form an interaction network. Clusters are created in the interaction network and the essential genes are detected using centrality measures.

The paper begin with literature survey, followed by methodology. The next session is result and discussion. The last part is conclusion and references.

2. Literature Review

The literatures on architecture of molecular networks reveal the cellular organization. The similarity between proteins can be evaluated in terms of their sequence or semantics. BLAST scores are used to evaluate sequence similarity [8]. The functionality of a gene is the semantic and it is expressed using GO terms. The method proposed by Wang et al. encode the biological meaning of a GO term into numeric value. It aggregates the semantic contributions of ancestor terms in directed acyclic graph [9-10].

The common similarity measures for comparing the annotation include cosine similarity, pairwise similarity, Jaccard's similarity and Levenshtein, based on distance and ratio measures [11]. Few others includeLatent Semantic Indexing (LSI), Word Mover's Distance and Latent Dirichlet Allocation (LDA) [12-13]. LSI is commonly used for web mining or natural language processing application and it suffer singular value decomposition. The main drawback of LDA is the linear growth in the number of parameters.

From the analysis of expression data sets it is clear that clusters can be formed from biological function. Lord et al., proposed a measure for annotation similarities in knowledge content between entries [14]. This can be used to perform similarity measure in an analogous manner over sequences. The validity of semantic similarity with sequence similarity is also evaluated [15]. The PPI network with the hub node is considered as a cluster and specifying the biological function of hub node [16].
Tiantian et al., proposed graph clustering algorithm, *EGCPI* takes two factors into consideration mainly topology of the network and attributes or features of interacting proteins [17]. Few researchers identified that centrality measures through subgraph are more efficient than the classic methods to find centrality [18]. Few researchers evaluated centrality measures by combining network topology and GO information for identifying essential proteins. Node centrality plays an important role in graph applications and biological network analysis [19].

Literatures indicate that most of the measures considers either lowest common ancestor or most informative common ancestor [20]. So it is highly essential to detect an effective similarity measure. In this research we perform similarity analysis between annotations. We consider a series of semantic similarity evaluations. The best similarity measure or the measure which indicate more similarity among the given annotation is considered. We follow the approach of an ensemble measure. PPI network is created among the genes to indicate the functionally similar genes. Using clustering, most essential node is detected using centrality measures.

3. Methodology

The methodology include the following steps
1. Similarity analysis under Jackknife methodology
2.Creating PPI network
3.Apply Centrality measures and use majority Voting Cluster Analysis

Similarity Analysis

In this step, the similarity between gene annotations is performed. To understand the best similarity technique to be applied on the GO data resource, few known similarity techniques were applied on the dataset. The measure with maximum similarity value is selected. In jackknife methodology, each method is considered and the best method is selected [21]. Algorithm 1 summarizes the general steps followed across all the similarity measures

Algorithm1: Similarity Analysis
Input: A set of genes and Annotation
Output: Similarity Matrix
: Identify the multiple similarity measures
2:Calculate the similarity between annotations
3: Generate similarity matrix for each measure
4: If the similarity index value>= .6, then
Similarity value = 1
else
Similarity value = 0
5: Count the positive values in each similarity matrix
6: Maximum positive matrix is elected.

Cosine Similarity

Cosine similarity measures can be used for measuring similar documents regardless of their size and is one of the common approaches used to find the similarity between the two documents [22]. Similarity between the two documents is identified based on counting

May-June 2020
ISSN: 0193-4120 Page No. 5124-5129

number of maximum words present as indicated in eq. (1) where A and B are annotations.

Cosine Similarity (A, B) = A. B = ||A|| ||B|| cos Θ (1)

Pairwise Document Similarity

Pairwise document similarity method is the technique based on terms in a document and the common terms (information) shared by two documents [23]. The weight of a term is assigned by a weighting scheme and indicates the significance of the term in that document.

Jaccard Similarity

Jaccard similarity is the ratio between the number of common words to the size of union of two words or sentences. We can find the similarity between two documents. If A and B are two objects, then similarity is calculated as given in eq (2).Lemmatization is performed to deduce words to the same root word.

$J(A,B) = |A \cap B| \div (|A| + |B| - |A \cap B|)$ (2)

Levenshte in Ratio and Distance Measure

Levenshtein distance with an unequal length, performs on strings. Levenshtein distance uses dynamic programming approach including string matching and spelling checking [11]. Levenshtein is a distance measure, but similarity can be calculated as in eq (3)

Similarity_Levenshtein(A,B)=1-Distance_Levenstein (A,B) (3)

Sequence Matcher using difflib

Sequence matchers can be used for comparing pairs of input sequences or strings. It focuses on comparing the longest contiguous matching subsequence between the two input sequences and finding the longest matching subsequence that contains no junk values. This is a class available in the python module named 'difflib'.

Creating PPI Network

The similar genes are linked to form a network in which nodes are the genes and the edges correspond to the similarity weight of 1. Based on the results obtained from multiple similarity measures, a threshold value is finalized. Any similarity score greater than or equal to .6 is considered as similar. Thus the two dimensional matrix obtained consist if a set of one and zero. The link is created between similar genes and they form a network [24].

Centrality Analysis

Centrality identifies the importance of a node or edge in the network within a graph. With respect to the *PPI* network, we apply centrality measures to identify the hub or influential node. Identification of hub node is important because of its strong influence over all other nodes. We consider four approaches were used for analyzing the network of gene interaction.

Closeness Centrality

Closeness centrality is calculated as the sum of the length of the shortest paths of a node to all other nodes. Closeness is calculated as given in eq (4).In equation, d(y,x) is the dissimilarity between vertices 'x' and 'y'.

$$C(x) = \frac{1}{\Sigma y\, d(y,x)} \quad (4)$$

Degree Centrality

The degree of vertex is the number of edges associated to a vertex, counted twice with loops. For a vertex 'v', for a graph G: = (V, E) with |V| vertices and |E| edges, centrality is defined as in eq. (5)

$$C_D(v) = \deg(v) \quad (5)$$

.

Betweenness Centrality

Betweenness centrality is based on the shortest paths. Ifσst is the number of shortest path from node *s* tot and $\sigma_{st}(v)$is the count of paths. Equation (6) represent this.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma st\,(v)}{\sigma st} \quad (6)$$

Eigenvector Centrality

Eigenvector centrality measures the influence of a nodes in the network. Scores are assigned to nodes in the network and higher scoring nodes are valuable in network.

Cluster Analysis

The centrality measured are calculated in PPI network. This network is now a cluster and the hub node represents the central node of the cluster. The cluster as a whole represents all the nodes with similar biological function. The network with three nodes or more than that was considered as the cluster.

4. Result and Discussion

The dataset used is from Gene Ontology database (www.geneontology.org) [25]. It consists of Electronic Annotations that are Swiss –Prot reviewed and manually curated Annotation or un reviewed Annotations.. The sample of dataset is given in table 1. This is implemented in python using scikit learn and pandas dataframe.

Under Jackknife methodology we implemented 5 different similarity measures. We considered cosine similarity, pairwise document similarity, Jaccard similarity, Levenshtein Ratio and sequence matcher. The sample output obtained by Levenshtein similarity by considering few genes like BUB1B, CENPE, INCENP, CENPA, CCNA2, MAD2L1 and NEK2 is

May-June 2020
ISSN: 0193-4120 Page No. 5124-5129

represented in fig 1. Here the similar genes (row and column headers) have a value 1 and dissimilar have a value 0 at their intersection. The similarity measures obtained with sequence matcher and Levenshtein are found to be same and it is finalized with 60% as threshold.

The result of different centrality measures are analyses and it is found that the same hub node is obtained with majority of the measures. Fig2 and fig 3indicate the result of two centrality measures and the gene INCENP is identified as hub node. Fig 4 indicate the clusters formed

	BUB1B	CENPE	INCENP	CENPA	CCNA2	MAD2L1
BUB1B	1	0	0	1	0	0
CENPE	0	1	0	0	0	1
INCENP	0	0	1	0	0	0
CENPA	1	0	0	1	0	0
CCNA2	0	0	0	0	1	0
MAD2L1	0	1	0	0	0	1
NEK2	0	0	0	0	0	0

Figure 1: Leven shte in distance measure

Table 2 represents the enrichment analysis for the clusters. It represents the annotation cluster, representative annotation term that is the common function among the cluster nodes and the enrichment score stating the number of node sin the cluster. Any newgene/protein provided by Gene Ontology can be applied with the following analysis and then add it to the respective cluster. This analysis helps us to identify the other common functions of the gene/protein and to evaluate the cluster properties.

['Gene/Protein' : Centrality value]
'INCENP': 0.16666666666666666,
'CDCA8': 0.0,
'CENPN': 0.0,
'CDC20': 0.0,
'ANAPC10': 0.0

Figure 2: INCENP as hub node with highest centrality value for Betweenness.

Table 1: Sample dataset

GENE	FUNCTION
B5KUL2	oxidation-reduction process, FMN hydroxy acid dehydrogenase
BUB1B	Mitotic checkpoint serine/threonine-protein kinase BUB1 beta
CENPE	Centromere-associated protein E; Microtubule plus-end-directed kinetochore motor
INCENP	Inner centromere protein; Component of the chromosomal passenger complex.

['Gene/Protein': Centrality value]
'INCENP':0.7071067811066628,
'CDCA8':0.49999999994351296,
'CENPN':0.49999999994351296,
'CDC20':1.0628924235733579e-05,
'ANAPC10': 1.0628924235733579e-05

Figure 3: INCENP as hub node with highest centrality value for Eigen value centrality.

5. Conclusion

Gene annotation represents the functional information about the gene. The correlation of semantic similarity to sequence similarity helps to predict protein function. We perform jackknife methodology for similarity evaluation. We analyzed semantic similarity between GO annotation and the optimum similarity measure is finalized. This helps to determine functionally similar genes. Furthermore PPI network created of similar genes identify the hub nodes of the network and can also detect the protein complexes by using various centrality measures. Therefore, this analysis can be used in bioinformatics to categorize genes/proteins into functionally similar clusters. Based on characteristics of the identified cluster group, further behavior of the gene can be predicted. The main limitation of any ontology approach is that incomplete GO annotation cannot be used to cover any statistical information.

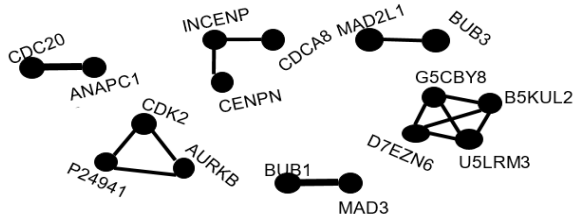

Figure 4: Clusters obtained

May-June 2020
ISSN: 0193-4120 Page No. 5124-5129

Table 2: Enrichment Analysis for clusters.

Annotation cluster	Representative Annotation Term	Enrichment Score
1 [P24941, CDK2,AURKB]	Cyclin-dependent kinase,Serine/threonine-protein kinase involved in the control of the cell cycle	3
2 [CENPN,INCENP,CDCA8]	Component of the chromosomal passenger complex, a complex that acts as a key regulator of mitosis	3
3 [D7EZN6,U5LRM3,G5CBY8,B5KUL2]	Catalytic activity, FMN hydroxy acid dehydrogenase domain-containing protein	4

References

[1] Wei Zhang , Jia Xu, Yuanyuan Li, and Xiufen Zou, “Detecting Essential Proteins Based on Network Topology, Gene Expression Data, and Gene Ontology Information”, IEEE Trans. On Comp. Bio. and Bioinfo., vol. 15, no.1, pp.109-116, Jan/Feb.2018.

[2] Victor Spirin and Leonid A. Mirny, “Protein Complexes and Functional Modules in Molecular Networks”, Science, vol.100, no. (21), pp. 12123-12128, Oct. 2003.

[3] Couto, F. M., & Silva, M. J. (2011). Disjunctive shared information between ontology concepts: application to Gene Ontology. Journal of biomedical semantics, 2(1), 5. http://www.jbiomedsem.com/content/2/1/5

[4] Avellaneda, M. J., Koers, E. J., Minde, D. P., Sunderlikova, V., & Tans, S. J. (2020). Simultaneous sensing and imaging of individual biomolecular complexes enabled by modular DNA–protein coupling. Imaging, 16, 20.

[5] A. Ali, R. Vishwanath, S. S. Patil, R. Abdulkader (2019).Alignment of Protein interaction network and disease prediction: A Survey. *International Journal of Advanced Trends in Computer Science and Engineering*, vol.8, No.4.

[6] Mazandu, G. K., Chimusa, E. R., & Mulder, N. J. (2017). Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. *Briefings in bioinformatics*, *18*(5), 886-901.

[7] A. Ali, R. Viswanath, S. S. Patil and K. R. Venugopal, "A review of aligners for protein protein interaction networks," *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore, 2017, pp. 1651-1655.

[8] J. ZHANG, S. Kwong and K. Wong, "ToBio: Global Pathway Similarity Search Based on Topological and Biological Features," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 336-349, 1 Jan.-Feb. 2019.

[9] H. Wang, Y. Du, J. Yi, Y. Sun and F. Liang, "A New Method for Measuring Topological Structure Similarity beween Complex Trajectories," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1836-1848, 1 Oct. 2019.

[10] D. Chicco and M. Masseroli, "Ontology-Based Prediction and Prioritization of Gene Functional Annotations," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 248-260, 1 March-April 2016.

[11] S. Zhang, Y. Hu and G. Bian, "Research on string similarity algorithm based on Levenshtein Distance," *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2017, pp. 2247-2251.

[12] P. van der Spek, S. Klusener and P. van de Laar, "Towards Recovering Architectural Concepts Using Latent Semantic Indexing," *2008 12th European Conference on Software Maintenance and Reengineering*, Athens, 2008, pp. 253-257.

[13] J. Yao, Y. Wang, Y. Zhang, J. Sun and J. Zhou, "Joint Latent Dirichlet Allocation for Social Tags," in *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 224-237, Jan. 2018.

[14] P.W. Lord, R.D. Stevens, A.Brass and C.A. Goble,“Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship Between Sequence and Annotation”, Science, Bioinformatics, vol. 19, no. 10, pp. 1275–1283, 2003.

[15] M. Siami, S. Bolouki, B. Bamieh and N. Motee, "Centrality Measures in Linear Consensus Networks With Structured Network Uncertainties," in *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 924-934, Sept. 2018.

[16] S. S. Bhowmick and B. S. Seah, "Clustering and Summarizing Protein-Protein Interaction Networks: A Survey," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 638-658, 1 March 2016.

[17] Tiantian He, and Keith C.C. Chan,“Evolutionary Graph Clustering for Protein Complex Identification”, IEEE Trans. on Comp. Bio. and Bioinfo., vol.15 ,

no.3, pp.892-904,May-Jun.2018.

[18] G. N. Vilarinho and E. E. Seron Ruiz, "Global Centrality Measures in Word Graphs for Twitter Sentiment Analysis," *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, Sao Paulo, 2018, pp. 55-60.

[19] Jalili, M., Gebhardt, T., Wolkenhauer, O., & Salehzadeh-Yazdi, A. (2018). Unveiling network-based functional features through integration of gene expression into protein networks. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1864*(6), 2349-2359.

[20] Acharya, Sudipta, Laizhong Cui, and Yi Pan. "Automated Hub-Protein Detection via a New Fused Similarity Measure-Based Multi-objective Clustering Framework." *International Symposium on Bioinformatics Research and Applications*. Springer, Cham, 2019.

[21] A. G. Holman, P. J. Davis, J. M. Foster, C. K. S. Carlow, and S. Kumar, "Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia malayi," BMC Microbiology, vol. 9, Nov. 2009, Art. no. 243, doi: 10.1186/1471-2180-9-243.

[22] A. F. Rojas Hernandez and N. Y. Gelvez Garcia, "Distributed processing using cosine similarity for mapping Big Data in Hadoop," in *IEEE Latin America Transactions*, vol. 14, no. 6, pp. 2857-2861, June 2016.

[23] Oghbaie, Marzieh, and Morteza Mohammadi Zanjireh. "Pairwise document similarity measure based on present term set." *Journal of Big Data* 5.1 (2018): 52.

[24] S. Bandyopadhyay and K. Mallick, "A New Path Based Hybrid Measure for Gene Ontology Similarity," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 116-127, Jan.-Feb. 2014.

[25] Harris, M. A., Clark, J., Ireland, A., et al. Gene Ontology Consortium 2004. *The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res*, *32*, D258-D261.DOI: 10.1093/nar/gkh036