EVALUATION OF SEMANTIC SIMILARITY BETWEEN GENE ONTOLOGY BASED ON PROTEIN FAMILY AND PATHWAY ANALYSIS

Anooja Ali[1], Ambili PS[2]
[1]*School of CSE, REVA University, Bangalore, India*
[1]*School of CSA, REVA University, Bangalore, India*

Article DOI: https://doi.org/10.36713/epra16626
DOI No: 10.36713/epra16626

ABSTRACT

An evolving vocabulary that explains the roles of proteins and genes is called gene ontology, or GO. Gene ontology (GO) describes the molecular, cellular, and biological levels of gene functioning. Semantic similarity gained relevance due to the widespread usage of gene annotations. There are a number of semantic similarity metrics that are available in the literature that concentrate on various strategies: distance-based techniques at the word level or gene product level, external documents, topology-based approaches that focus on boundaries, ancestor or child nodes. We presume that combining all of these element's results in a methodical way to gauge the degree of similarity across GO annotation items. We have conducted a detailed analysis of the biological pathways and GO keywords, and we have created a semantic measure of similarity called SimGOT. SimGOT takes into account topology-based similarity measures, membership of words in fuzzy clustering, and semantics hidden in the ontology or information content of a term. UniProt is used to build the datasets that are positive and negative. We compared four existing GO-based semantic similarity metrics based on semantic similarity, Pearson's correlation coefficient, and Protein Family (Pfam) subdomain group similarity. The superiority of SimGOT over alternative semantic similarity metrics is demonstrated by the experimental findings.

KEYWORDS—*Gene Ontology, Pearson's correlation coefficient, Protein family, SimGOT, UniProt*

I. INTRODUCTION

The regulation of cellular life is significantly influenced by proteins. They provide the key to understand the requirements to support life. Thus proteins are requisite to live [1]. Hundreds to thousands of amino acids together form a protein molecule. Bioinformatics facilitates the statistical analysis of protein sequences thereby annotating the genome to predict their structure and to understand the functionality [2]. Bioinformatics researchers often use similarity measures to compare one protein with another.

A sequence of amino acids constitutes one protein. These sequences correspond to sequence similarity. The similarity is commonly compared with the BLAST algorithm and the BLOSUM62 scoring matrix [3]. *GO* is an adaptable database containing the gene functions of various organisms like animal, plant, human and microbial genomes. GO consortium updates the database on regular basis. The three taxonomies that enclose the pertaining biological knowledge are Molecular Function (*MF*), Biological and Cellular Components (BP and *CC*). The ontology for these taxonomies contains several processes, related to each other and referred to as *GO* terms. The *GO* term for each taxonomy (*MF*, *BP* and *CC*) can be downloaded from the ontology website.[4].

The relationship between the pairs in GO is 'part-of' and 'is-a'. Few characteristics are inherited from ancestors. So GO is a Directed Acyclic Graph (DAG). Each term in GO indicates the role of the protein in performing MF, BP and CC processes. The hierarchical relationship is represented by edges in a DAG, whereas nodes stand in for GO words.

GO helps to predict essential proteins. Information content present in an annotation can be used to measure semantic similarity between proteins. Genes are semantically similar if they have interconnected MF, BP and CC functionalities. Consider a subgraph of CDC20 gene (Cell Division Cycle 20), a protein-coding gene with GO: 1990333 the mitotic checkpoint complex [5]. Fig 1 indicates the ancestor chart for CDC20. CDC20 is a pre-invasive hub gene for cervical cancer.

Semantic similarity is the similarity score of ontology terms between interacting proteins. Semantic similarity analysis can uncover protein clustering [6], pathway modeling [7] and protein interaction predictions [8]. Functional similarities are advantageous for various applications hence; it is required to ensure that the similarity measures are reliable. Correlation between sequence and semantic similarity may not be existing with all protein pairs. Pearson's association Coefficient is a frequently used metric to determine this association. [9]. While experimenting with the existing semantic and sequential similarity measures, we identified a few questions that need to be addressed. When the correlation between the similarity measures is low, each similarity measure becomes independent. When correlation is high, can query based on semantic similarity be an alternate for the existing sequence matching methodologies. Thus a systematic method is required to investigate this correlation.

Most of the current methods don't take into account all of the important GO graph topological properties. In order to extract the proteins that the present tools are missing, it also advocates using semantic tools including improving the ones that are already in place. Semantic similarity is measured by most information content-based methodologies using the information content between GO keywords.

In this paper, we attempt to resolve these issues by inquiring about the relationship between various similarity measures and the validity of the proposed method is compared with the existing similarity measures. SimGOT incorporates every significant similarity computation approach. The following is a list of the planned work's primary contributions:

1. A new topology based similarity measure, struct_depth().
2. Multifact_sim() is a multi-factored similarity measure that incorporates weight function, participation for every term, and fuzzy clustering.

The remainder of the document is structured as follows: A review of the relevant literature opens the following section.

II. RELATED WORK

There are several methodologies available in the literature for calculating semantic similarity based on GO terms and topology. In this section, we present a systematic study of the various methods available in the literature.

Approaches to semantic similarity may be roughly categorised as Node-based, Edge related, Hybrid and Node-based methods consider the features of GO terms which are linked to their parent or child and they often query the nodes [10]. The information quality among GO words is taken into consideration by very few node-based techniques. In line with the information content principle, if a GO term is t, the probability to detect the child of t is P(t). The information content present in the term t can be denoted by – log P(t) [11]. Eq.1 denotes this.

$$IC(t) = -\log P(t) \qquad (1)$$

According to this, if the frequency of usage of the GO term becomes common in a specific database, then the GO term is considered less informative. Edge-based methods are dependent on distance function either based on shortest path or common path to an ancestor in DAG [12]. If two GO terms have a common ancestor, then the semantic similarity between them is measured with the concept of information content, either as selecting the Most Informative Common Ancestor (MICA) or Common Disjunctive Ancestor (CDA) [13].

Similarity measure proposed by Resnik, Lin [14], Jiang [15] selects the MICA. The advantage of Lin and Jiang's measure over Resnik is the normalization of value from 0 to 1. Lin normalized the similarity by averaging the information content of two terms [16]. Thus the similarity measure considers the information available in query terms. Similarity by Lin's measure from 0 to 1. S (t[i], t[j]) denotes the set of parents shared by terms t[i] and t[j]. Resnik does not consider the distance of the LCA (Lowest Common Ancestor). So, if two terms have a common ancestor and if they are at different levels of GO, still their semantic similarity remains the same. The distance of the LCA is taken into account by Lin's similarity measure, but the depth of the common ancestor is not. Few researchers even combined Resnik's, Lin's and Jiang's similarities. Schlicker considered the annotation probability of the ancestor with more information content [17].

A common drawback with these methods is they consider only MICA and not the CDA. Wang's similarity measure considers topological information and ignores annotation. Wang's method had a substantial advantage over other information content-based methods. Nagar and Al- Mubaid proposed a hybrid measure that uses the shortest path based on topology and information content from DAG. Depending solely on correlation and predicting protein functionality may produce errors [18]. While group-wise approaches solely assess functional similarity, pairwise methods assess the semantic similarity of GO keywords. A strong association between sequencing and annotation similarity was investigated by Lord et al [19].

According to the literature review, there is no set method for determining the optimal similarity metric. Few of the approaches ignore the CDA and consider only MICA. Few approaches do not consider any topological feature of the graph. All the existing measures used the different properties of GO term and they are auxiliary to each other. Combining the GO term and topology are delimited in literature. In this research, we propose SimGOT, to estimate similarity at pairwise or GO term level and at the topological level.

III. METHODOLOGY

The three main phases of SimGOT are as follows:

1. The number of connections of a node with smooth information is taken into consideration by the suggested technique, which employs depth as a factor for similarity measure.
2. A review is conducted of the information content found on the shortest path connecting the cluster centre and the GO term.
3. Fuzzy clustering allows a GO term to be a member of more than one clustering. Unlike other clustering algorithms, clusters can have overlapped members.

It is important to mention *GO* modernization. *GO* structure is updated frequently with the emergence of new annotations and the relationship between the annotation and path. Subsequently, the *GO* database will be updated regularly. So, the features of *GO* terms like annotation, path, depth, information content are falsified.

A. Similarity based on struct_depth

Let t be a GO term. The number of ascendants or descendants that are either directly or indirectly linked to t in an ontology is indicated by the symbol N(t). Depth of a word is defined as the ratio of this to the number of GO terms associated with a certain ontology and is represented as depth (t). The corpus is |O| in size.

A combination of information content and a topological metric, as shown in equation (2).

$$depth\ (t) = \frac{N(t)}{|O|} \qquad (2)$$

Each node in the graph is associated with *depth (t)* indicating connectors of the node. Nodes with *depth (t)* above the threshold are elected as cluster centers by dividing with the height of the tree as in eq. (3)

$$Struct\_depth(t) = \frac{depth(t)}{depth(GO)-1} \qquad (3)$$

The node with the greatest connection level must be the cluster centre. As with Lin's and Jiang's measure, relying just on LCA will not be adequate to identify the cluster centre and will result in a shallow annotation problem. Genes with shallow hierarchical annotations have high similarity.

To evaluate the similarity between terms, t1 and t2, the average struct_depth of both terms are calculated. The proposed method to detect cluster center consider all the features including the number of interconnected nodes, number of GO term and depth of GO tree. The proposed method comprehends fuzzy clustering because each term can belong to more than one cluster.

B. Similarity based on multifact_sim

Few researchers considered the distance of the shortest path from the cluster center to every other node in the network. The advantage of this method over other existing methods is the combination of information content along with path length. Let P_1 indicate the path from the cluster center, *c* to a node t_1 in the network and P_2 indicates the path from *c* to the node t_2.

The difference between the two terms t_1 and t_2 concerning the cluster head is detected. Let C_1 and C_2 be the cluster center of t_1 and t_2 respectively. The membership function of t_1 with the cluster C_1 be m($t_{1\text{-}}$ C_1) and the function of t_1 with the cluster C_2 be m(t_1- C_2). The membership function of t_2 with the cluster C_1 be m($t_{2\text{-}}$ C_1) and t_2 with cluster C_2 be m($t_{2\text{-}}$ C_2).

The difference between the terms concerning cluster C_1 for t_1 and t_2 is indicated as [*m*($t_{1\text{-}}$ C_1)- *m*(t_2- C_1)]. Similarly, for cluster C_2 the difference between the membership function is indicated as [*m*($t_{1\text{-}}$ C_2)- *m*(t_2- C_2)]. These differences are represented by *Diff* (C_1) and *Diff* (C_2). If multiple similarity measures are considered, then the similarity measure with maximum similar candidates are referred to in the next step. Following this principle, the maximum difference is considered as in eq (4).

$$MaxDiff(t1, t2) = Max[Diff(c1), Diff(c2)] \quad (4)$$

The precision of the proximity metric being utilised determines how effective any similarity measure will be. Only interacting proteins will have strong semantic similarity as determined by GO keywords . Our suggested technique does this by combining the topology with the term's information content.

The semantic similarity for protein pairings is computed using Best Match Average (BMA). BMA performs better biologically than average and maximum methods. Average or maximum use is restricted to the given application. (5) provides the suggested similarity equation.

$$X = weight(t_1, t_2) + struc\_depth(t_1, t_2) \qquad (5)$$

IV. RESULTS AND DISCUSSION

Benabderrahmane et al. employ a benchmark dataset to assess SimGOT in order to assess different GO features. SimGOT outperforms other cutting edge methods in terms of correlation and Pfam similarity. We generated a list of positive and negative interactions by analysing the UniProt dataset. The Pfam score is determined by dividing the total number of families that proteins share by the number of protein families that they share . Under BMA, resemblance scores are shown. We employ Nunivers for normalisation and the GO universal measure, so BMA may be used to determine functional similarity at the end. The correlation between sequencing and semantic similarity is determined using Pearson's correlation coefficient. BLAST log bit score is used to calculate sequence similarity .

Evaluation is carried out using GO:0003674 as the DAG. We conducted an evaluation based on the MF ontology with 27 direct descendants of this GO word. The similarity between the GO word pairs GO:0046572 and GO:0016829, GO:0060089 and GO:0004872 is displayed in Table 1. Using information content-based methodologies such as Resnik, Lin, and Wang, we assessed SimGOT.

Pfam clans' intraset similarity is computed. The evaluation is conducted using the dataset from genes found in the same clan have comparable molecular functions, and MF ontology is used to access similarities [20]. The clans utilised in the similarity study are listed in Table 2. The Pearson correlation coefficients for the three ontologies are shown in Table 3. SimGOT consider all the ancestors shared between the terms. Considering MICA or CDA alone will not be appealing for a denser graph because the information content of some useful ancestors will not be considered.

Table 1: Semantic similarity comparison of *SimGOT* with other Information Content methods (Resnik, Lin, Wang, *GOGO*) for the *GO* term pairs (GO: 0046572 and GO: 0016829) and (GO:0060089 and GO:0004872).

Approach	Similarity (0046572,0016829)	Similarity (0060089, 0004872)
Resnik	0.082	0.311
Lin	0.135	0.762
Wang	0.610	0.715
GOGO	0.376	0.534
SimGOT	**0.396**	**0.622**

Table 2: The Pfam clans that were utilised to determine each clan's gene count and degree of similarity.

T*Pfam* Clan	No: of genes
ALDH	15
BIR	8
FBD	7
Flavo-protein	8
6PGDC	7

Table 3: A comparison of the CC, BP, and MF ontologies' Pearson Correlation Coefficients. The values with the highest values are bolded.

Approach	*CC*	*BP*	*MF*
Lord	0.523	0.521	0.625
Al Mubaid	0.514	0.492	0.543
Wang	0.637	0.532	0.622
TopoICSim	0.6346	0.528	0.623
SimGOT	**0.644**	**0.613**	**0.704**

Table 4: The Pearson Correlation Coefficient for MF, CC, and BP ontologies on IEA- and IEA+ between sequence and similarity scores. The ontologies with the highest values are indicated.

Approach	**Pearson's Correlation for *IEA-***			**Pearson's Correlation for *IEA+***		
	MF	***CC***	***BP***	***MF***	***CC***	***BP***
Lord	0.529	0.428	0.411	0.562	0.416	0.511
Al Mubaid	0.513	0.426	0.422	0.540	0.422	0.531
Wang	0.522	0.431	0.416	0.532	0.428	**0.549**
TopoICSim	0.521	0.431	**0.421**	0.540	0.436	0.512
SimGOT	**0.518**	**0.443**	0.438	**0.531**	**0.448**	0.509

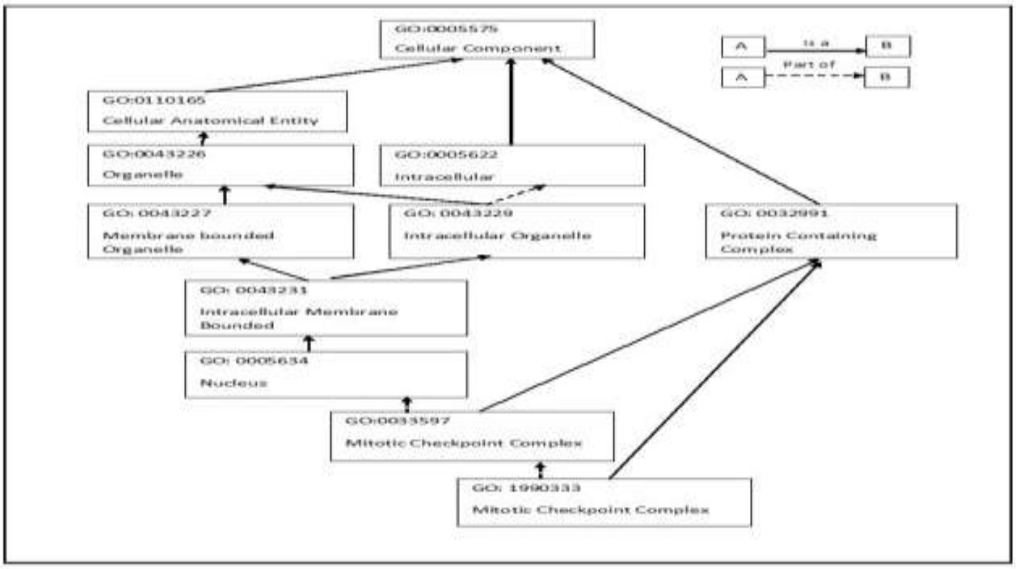

Fig 1 Ancestral chart for CDC-20, mitotic checkpoint complex with GO- 199033

SimGOT exhibits superior performance over the topology-based approach by Wang et al. The main advantage of *SimGOT* is the apprehension of fuzzy clustering for *GO*, shortest path distance to the cluster center, membership. *SimGOT* excel over other information content-based techniques based on *MICA* and *CDA* approaches. We combine information and topology aspects of the term and estimate the similarity between them.

A. Evaluation with Interacting Dataset

Dataset used is from the Gene Ontology database [21]. An optimum coverage can be obtained only by including *IEA+* and *IEA-*. Protein interactions reviewed in UniProt are considered as the positive dataset [22]. This constitutes 3,500 interactions. A negative dataset is created by considering un reviewed annotations from UniProt. The sequence and semantic similarity are correlated to each other because the standard deviation between them is ±2SD. So, Pearson correlation coefficient can be used for further calculation.

When it comes to MF and CC ontologies for the IEA+ and IEA-datasets, SimGOT shows the strongest association. Table 4 illustrates this. This is because multifactor similarity was taken into account. Terms that correspond to many clusters are taken into account via fuzzy clustering. The depth of the term plus best match average increase SimGOT's efficiency. When all three-ontology ontology are evaluated, the MF ontology shows a stronger association.

V. CONCLUSION

In this research, we provide an improved method for evaluating semantic similarity for GO words, called SimGOT, which is based on the term's information content as well as topological variables of DAG, such as the terms' structure depth, membership, size, and shortest path. We test SimGOT on the Pfam clan based on intraset similarity. SimGOT exhibited improved performance by considering the benchmark datasets. SimGOT shows robust performance over existing approaches like Lord, Wang, Al Mubaid and TopoICSim. We evaluated the pairwise gene similarity and compared it with other information content-based approaches. We also assess the performance of SimGOT for Pearson's correlation among sequence and semantic similarity on positive and negative datasets and evaluated that. The research on considering all the ancestors or particular ancestors is still in progress.

VI. REFERENCES

1. *Ramachandra, H. V., et al. "An Optimization on Bicluster Algorithm for Gene Expression Data." 2023 4th IEEE Global Conference for Advancement in Technology (GCAT). IEEE, 2023.*
2. *Ali, Anooja, Vishwanath R. Hulipalled, and S. S. Patil. "Centrality measure analysis on protein interaction networks." 2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET). IEEE, 2020. DOI: https://ieeexplore.ieee.org/abstract/document/9557447*
3. *Patil, S. S., Anooja Ali, and A. Ajil. "Approaches for network analysis in protein interaction network." International Journal of Human Computations & Intelligence 2.2 (2023): 47-54. DOI: https://doi.org/10.5281/zenodo.7900226*
4. *GO-Consortium (2009). The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Research 38:D331–D335.*
5. *Ali, Anooja, et al. "Detection of gene ontology clusters using biclustering algorithms." SN Computer Science 4.3 (2023): 217. DOI: https://link.springer.com/article/10.1007/s42979-022-01624-w*
6. *Dutta, P., Basu, S., & Kundu, M. (2017). Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph. IEEE/ACM transactions on computational biology and bioinformatics, 15(3), 839-849. DOI: https://doi.org/10.1109/TCBB.2017.2689762*
7. *H. V. Ramachandra, A. Ali, P. S. Ambili, S. Thota and P. N. Asha, "An Optimization on Bicluster Algorithm for Gene Expression Data," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/GCAT59970.2023.10353373.*
8. *Ali, A., Viswanath, R., Patil, S. S.,et al. (2017). A review of aligners for protein protein interaction networks. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1651-1655). IEEE. DOI: https://doi.org/10.1109/RTEICT.2017.8256879*
9. *Gogtay, N. J., & Thatte, U. M. (2017). Principles of Correlation Analysis. The Journal of the Association of Physicians of India, 65(3), 78–81. PMID: 28462548.*
10. *Zhang, J., Jia, K., Jia, J., et al. (2018). An improved approach to infer protein-protein interaction based on a hierarchical vector space model. BMC bioinformatics, 19(1), 161. DOI: https://doi.org/10.1186/s12859-018-2152-z*
11. *Pillai, A.P.S. (2023), "AIoMT-Assisted Telemedicine A Case Study of eSanjeevani Telemedicine Service in India", Handbook of Security and Privacy of AI-Enabled Healthcare Systems and Internet of Medical Things, pp. 445–464, DOI: 10.1201/9781003370321-19.*
12. *Köhler, S., Vasilevsky, N. A., Engelstad, M., et al. (2017). The human phenotype ontology in 2017. Nucleic acids research, 45(D1), D865-D876. DOI:10.1093/nar/gkw1039*
13. *Couto, Francisco M., Mário J. Silva, and Pedro M. Coutinho. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. Proceedings of the 14th ACM international conference on Information and knowledge management. 2005.*
14. *Ambili P S, Agnesh L, & Arun K V. (2023), "Siamese Neural Network Model for Recognizing Optically Processed Devanagari Hindi Script", International Journal of Computational Learning & Intelligence, 2(3), 107–113, https://doi.org/10.5281/zenodo.8210372.*
15. *Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008. https://arxiv.org/pdf/cmp-lg/9709008.pdf*
16. *Ramachandra, H. V., et al. "Ensemble machine learning techniques for pancreatic cancer detection." 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC). IEEE, 2023. DOI: https://ieeexplore.ieee.org/ abstract/document/10200380*
17. *Schlicker, A., Domingues, F. S., Rahnenführer, J., et al. (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics, 7(1), 302. DOI: 10.1186/ 1471-2105-7-302*
18. *Lin, D. (1998, July). An information-theoretic definition of similarity. In Icml (Vol. 98, No. 1998, pp. 296-304.*
19. *Lord, P. W., Stevens, R. D., Brass, A., et al. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19(10), 1275-1283. DOI: https://doi.org/10.1093/ bioinformatics/ btg153*
20. *A Ali, VR Hulipalled, SS Patil and RA Kappaparambil, "DPCCG-EJA: detection of key pathways and cervical cancer related genes using enhanced Johnson's algorithm", Int J Adv Sci Technol, vol. 28, no. 1, pp. 124-138, 2019.*
21. *Jain, A., Perisa, D., Fliedner, F., von Haeseler, A., & Ebersberger, I. (2019). The evolutionary traceability of a protein. Genome biology and evolution, 11(2), 531-545. DOI: https://doi.org/10.1093/gbe/evz008*
22. *Ali, Anooja, et al. "Pareto Optimization Technique for Protein Motif Detection in Genomic Data Set." International Conference on Information, Communication and Computing Technology. Singapore: Springer Nature Singapore, 2023. https://doi.org/10.1007/978-981-99-5166-6_65*