SEVAS
Publishing

Computing Technology Research Journal
Comput. Technol. Res. J. (Oct, 2022) 1(1):1-4

ISSN: 2583-6501
DOI: 10.58973/CTRJ.22111
@ All Rights Reserved

A NOVEL SEMANTIC SIMILARITY SCORE FOR PROTEIN DATA ANALYSIS

Anooja Ali*
School of CSE, REVA University, Bengaluru
*Email: anoojaali@gmail.com

Vishwanath R Hulipalled
School of C&IT, REVA University, Bengaluru

S.S. Patil
Department of Agricultural Statistics, University of Agriculture Sciences, Bengaluru

Received: 10th September 2022, Accepted: 3rd October 2022 and Published: 3rd October 2022

ABSTRACT

Aim: A similarity evaluation measure for Gene Ontology GO terms is developed.
Results: The proposed method takes into account the semantics hidden in ontologies or the term level information content, membership of term, and topology-based similarity measures. The proposed method is evaluated on positive and negative dataset of UniProt, Protein family clans and the Pearson's correlation with other existing methods.
Conclusion: The experimental results exhibited a major supremacy of the proposed method over other semantic similarity measures.

Keywords: Annotation, Information Content, Membership, Topology

HIGHLIGHTS:

1. **An improved approach for semantic similarity evaluation for GO terms based on the information content and the topological factors is developed.**
2. **The proposed method shows highest correlation for MF (Molecular Function) ontology.**

INTRODUCTION

The proposed method combines all the important techniques of similarity calculation including a multi-factored similarity measure that combines fuzzy clustering, Information content membership of each term and weight function.

The three major steps involved in similarity evaluation measure for Gene Ontology (GO) terms are listed below.

1. The proposed method uses depth as a factor for similarity measure by considering the number of children of a node with fine-grained information content.
2. Information content present along the shortest path from the GO terms to Many Integrated Core Architecture (MICA) is evaluated.
3. Membership of a term is evaluated with Gaussian membership function.

Evaluating the Depth of a Term

Consider a *GO* term, *t*. *N(t)* denotes the number of descendants linked to *t* either directly or indirectly in an ontology. The proportion of this with the number of *GO* term corresponding to a particular ontology is expressed as the depth of a term, denoted as *depth (t)*. The size of the corpus is |O|. A topological measure combined with information content and this is represented in eq. (2).

$$Depth\ (t) = -ln\frac{N(t)}{|O|} \qquad 2$$

Semantic Similarity Score

The cluster center needs to be the highest connected node. Considering LCA (Lowest Common Ancestor) alone will not be sufficient to detect the cluster center and it leads to shallow annotation problem as in the case with Lin's and Jiang's measure. Genes annotated at shallow levels of the hierarchy result in high similarity.

Similarity based on shortest Path

The proposed method considers multiple factors for similarity evaluation and these factors are mutually exclusive. We consider similarity based on shared information content, shortest path or topology-based, membership of terms.

Detecting the shortest path from the GO term to MICA

The proposed method that consider the distance of the shortest path between the term to MICA [1]. Considering the ancestor that holds the majority information is more worth than considering all the ancestors. This is du eto the fact that a node can be a parent for multiple child node. The shortest path is between the terms t1 and t2 is calculated as in eq. (3)

$$dist(t_1, t_2) = \frac{\sum_{t_1 \text{€} path_1} \frac{1}{IC[t_1]} + \sum_{t_2 \text{€} path_2} \frac{1}{IC[t_2]}}{2} \qquad 3$$

Best Match Average (BMA) calculates the semantic similarity for protein pairs [2]. From a biological aspect, BMA outperforms average and maximum approaches. The use of average and maximum is bound to the application. We derived a unified framework for semantic similarity calculation by combining annotation, weight function, and topology approaches.

Evaluating the membership of the term

The membership of the term is evaluated using Gaussian membership function, All the default parameters are used in membership evaluation and is calculated as in eq (4).

$$M(t) = exp\left\lfloor\frac{-(t-m)^2}{2k^2}\right\rfloor \qquad 4$$

The efficiency of any similarity measure depends on the accuracy of the proximity measure used. Semantic similarity calculated by GO terms will be high only for interacting proteins. This is swamped in our proposed method by incorporating information content of the term and the topology.

Semantic Similarity Score

Best Match Average (BMA) calculates the semantic similarity for protein pairs [2]. From a biological aspect, BMA outperforms average and maximum approaches. The use of average and maximum is bound to the application.We derived a unified framework for semantic similarity calculation by combining annotation, weight function, and topology approaches.

RESULTS:

The proposed method is evaluated on a benchmark dataset for evaluating various GO properties [3]. It exhibits higher correlation and *Pfam* similarity against other state of art techniques. We used the UniProt dataset for analysis and created a positive and negative set of interactions [4]. *Pfam* score is calculated as the total count of families shared by the proteins to the count of protein families they share [5]. The similarity scores are provided under BMA. We follow GO universal measure, Nunivers for normalization and so BMA can be used to finalize functional similarity [6]. Pearson's correlation coefficient is used to find the correlation between sequence and semantic similarity. Sequence similarity is calculated by BLAST log bit score [7].

Intraset similarity and semantic similarity evaluation

Evaluation is performed with a *DAG* (Directed Acyclic Graph) of GO:0003674. We performed an *MF* (Molecular Function) ontology (announced on September 10, 2016) based evaluation. This *GO* term has 27 direct descendants. Table 1 shows the similarity between *GO* term pairs *GO*:0046572 and GO:0016829, GO:0060089 and *GO*:0004872. The performance is evaluated with information content-based approaches including Resnik, Lin, Wang, and *GOGO*. *GOGO* is a webserver that calculates the semantic similarity between *GO* terms [8]. It is clear from our observations that nodes with more number of children are negatively correlated to the information content of the *GO* term.

The intraset similarity is calculated of *Pfam* clans. The dataset used in [3] is used for evaluation. Genes present in the same clan share the molecular function and similarity is accessed based on MF (Molecular Function) ontology [9]. Table 2 indicates the clans used for similarity analysis. Figure 1 indicates the *Pfam* clan and the intraset similarities for *MF* annotations and Table 3 indicates the Pearson correlation coefficient for the three ontologies. The proposed method outperforms other approaches including Wang, Lord, TopoICSim and Al Mubaid measure.

The proposed method exhibits high correlation for MF and CC (Cellular Component) ontologies for IEA+ (with electronic annotations) and IEA- (without electronic annotations) datasets (Table 4). This is due to the consideration of multifactor similarity. Fuzzy clustering considers terms that belong to multiple clusters. Best match average along with a depth of the term improves the efficiency. All three ontologies are used for evaluation and MF ontology exhibits a higher correlation.

CONCLUSION

In this paper, we present an improved approach for semantic similarity evaluation for GO terms based on the information content of the term and the topological factors of DAG like structural depth and membership. It is important to note that information content-based approaches should not be biased on the corpus.

REFERENCES:

1. https://doi.org/10.1504/ijdmb.2014.062887
2. https://doi.org/10.1186/2041-1480-2-3
3. https://doi.org/10.1186/1471-2105-11-588
4. https://doi.org/10.1093/nar/gki070
5. https://doi.org/10.1093/nar/gkv1344
6. https://doi.org/10.1371/journal.pone.0113859
7. https://doi.org/10.1109/TCBB.2017.2695542
8. https://doi.org/10.1038/s41598-018-33219-y
9. https://doi.org/10.1186/s12859-016-1160-0

TABLES & FIGURES:

Table 1: Semantic similarity comparison of proposed method 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). *SimWOS* exhibits the highest similarity with the two gene pair.

Approach	Similarity(0046572,0016829)	Similarity(0060089, 0004872)
Resnik	0.081	0.310
Lin	0.134	0.760
Wang	0.612	0.712
GOGO	0.386	0.544
Proposed Method	**0.398**	**0.612**

Table 2: The *Pfam* clans used for similarity calculation and the number of genes existing in each clan.

T*Pfam* Clan	No: of genes
ALDH-like	18
BIR-like	9
FBD	6
Flavoprotein	7
6PGD_C	8

Table 3: Comparison of the Pearson Correlation Coefficient for CC, BP (Biological Process), and MF ontologies. The highest values are indicated in bold. The proposed method shows the highest correlation for all the three ontologies.

Approach	*CC*	*BP*	*MF*
Lord	0.513	0.511	0.615
Al Mubaid	0.504	0.490	0.553
Wang	0.627	0.512	0.612
TopoICSim	0.636	0.518	0.633
Proposed	**0.654**	**0.623**	**0.714**

Table 4: Pearson's Correlation Co-efficient between sequence and similarity scores for *MF*, *CC* and *BP* ontologies on IEA- and IEA+. The highest values for each ontology is highlighted

Approach	Pearson's Correlation for *IEA-*			Pearson's Correlation for *IEA+*		
	MF	*CC*	*BP*	*MF*	*CC*	*BP*
Lord	0.519	0.418	0.421	0.572	0.426	0.521
Al Mubaid	0.533	0.436	0.412	0.580	0.432	0.541
Wang	0.512	0.421	0.406	0.552	0.418	**0.559**
TopoICSim	0.501	0.441	**0.431**	0.560	0.446	0.502
Proposed Method	**0.538**	**0.453**	0.428	**0.591**	**0.438**	0.539

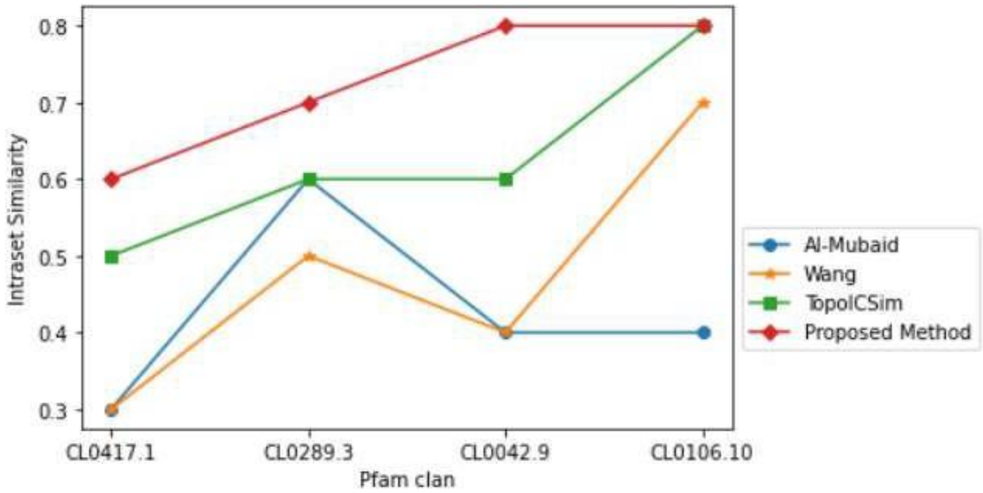

Figure 1: Comparison of *Pfam* similarity for *MF* ontology between the *Pfam* clans. The intraset similarity is estimated between all the genes present in the clan. The proposed method shows the highest *Pfam* similarity for *MF* ontology