

A NOVEL SEMANTIC SIMILARITY SCORE FOR PROTEIN DATA ANALYSIS

Anooja Ali*

School of CSE, REVA University, Bengaluru, 560064, India
*Email: anoojaali@gmail.com

Vishwanath R Hulipalled

School of C&IT, REVA University, Bengaluru, 560064, India

S.S. Patil

Department of Agricultural Statistics, University of Agriculture Sciences, Bengaluru, 560065, India

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

ABSTRACT

Background: Semantic similarity between Gene Ontology (GO) terms underpins numerous applications in genomics, yet existing measures typically rely on a single strategy — topology, information content, or node/edge structure — and no individual approach consistently outperforms the others.

Objective: This study develops a similarity evaluation measure for GO terms that jointly incorporates information content, term membership, and topology-based similarity.

Methods: The proposed method computes a topological depth factor from the number of descendants of a term, an information-content-based shortest-path distance to the Most Informative Common Ancestor, and a Gaussian membership function; these are combined via Best Match Average and evaluated on positive and negative UniProt interaction datasets and Pfam protein family clans.

Results: The proposed method achieved higher Pfam similarity and Pearson correlation than Resnik, Lin, Wang, GOGO, Lord, Al-Mubaid, and TopoICSim overall, with the strongest correlation observed for the Molecular Function (MF) ontology.

Conclusion: The experimental results demonstrate a clear advantage of the proposed method over existing semantic similarity measures for GO term and protein-clan analysis.

Keywords: Annotation, Information Content, Membership, Topology

HIGHLIGHTS:

1. An improved approach for semantic similarity evaluation of Gene Ontology (GO) terms is developed by combining information content and topological factors.
2. A multi-factor similarity measure integrating fuzzy clustering, term membership, and a weight function is proposed.
3. The proposed method achieves the highest Pfam similarity and the highest correlation for the Molecular Function (MF) ontology among all compared approaches.

INTRODUCTION

Understanding protein function is central to bioinformatics, supporting genome annotation and functional prediction. Gene Ontology (GO) describes gene functions across organisms through three taxonomies — Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) — organized as a Directed Acyclic Graph (DAG) of terms linked by ‘part-of’ and ‘is-a’ relationships. Existing semantic similarity approaches are broadly node-based, edge-based, or hybrid. Node-based methods, including those of Resnik and Lin, estimate similarity from the information content of the Most Informative Common Ancestor (MICA) but disregard the depth of the common ancestor [1,2]. Edge-based methods instead rely on shortest-path distance within the DAG; Wang’s measure incorporates topology but ignores annotation [3], and Lord et al. found a strong correlation between sequence and annotation similarity [4]. No single approach combines all the relevant topological features of the GO graph with term-level information content.

We therefore develop a similarity measure for GO terms combining information content, term membership, and topology. Its contributions are: (i) a multi-factored measure combining fuzzy clustering, information content membership, and a weight function; (ii) a depth-based topological factor from the number of children of a node; and (iii) a shortest-path information content measure to the MICA combined with a Gaussian membership

function. The method is evaluated against Resnik, Lin, Wang, GOGO, Lord, Al-Mubaid, and TopoICSim on GO term pairs, Pfam clans, and UniProt interaction datasets.

MATERIALS AND METHODS

The proposed method estimates semantic similarity for Gene Ontology terms in three stages: a topological measure based on the depth of a term, a shortest-path-based information content measure to the Most Informative Common Ancestor (MICA), and a membership measure computed with a Gaussian membership function. GO is treated as a Directed Acyclic Graph, in which each term inherits characteristics from its ancestors.

3.1 Evaluating the Depth of a Term

For a GO term t , the information content is defined using the standard formulation, where $P(t)$ is the probability of observing t or its descendants in the corpus:

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

Let $N(t)$ denote the number of descendants linked to term t , either directly or indirectly, within an ontology, and let $|O|$ denote the size of the corpus. The depth of a term, $depth(t)$, is expressed as the proportion of $N(t)$ to $|O|$; combining this topological measure with information content gives:

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

Relying on the Lowest Common Ancestor (LCA) alone is insufficient and leads to the shallow-annotation problem observed with Lin's and Jiang's measures, in which genes annotated at shallow levels of the hierarchy exhibit artificially high similarity.

3.2 Similarity Based on Shortest Path

The proposed method combines three mutually exclusive factors: shared information content, shortest-path distance, and term membership.

3.2.1 Detecting the Shortest Path from the GO Term to MICA

Rather than considering all ancestors, the proposed method uses the distance of the shortest path between a term and MICA [5], since the ancestor with the majority of the information content is more informative than the complete ancestor set. The shortest path between terms t_1 and t_2 is calculated as:

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

Best Match Average (BMA) derives protein-pair similarity from term-pair similarity [6], outperforming simple average and maximum aggregation.

3.2.2 Evaluating the Membership of the Term

The membership of a term is evaluated using a Gaussian membership function, with default parameters throughout:

$$M(t) = \exp \left[\frac{-(t-m)^2}{2k^2} \right] \quad (4)$$

Because semantic similarity is high only for genuinely interacting proteins, incorporating both information content and topology improves the reliability of the proximity measure.

3.3 Semantic Similarity Score

The proposed method derives a unified similarity score by combining the depth-based topological factor, the shortest-path information content, and the Gaussian membership factor through BMA, then normalizes it using the GO universal measure, Nunivers, to finalize the functional similarity score.

RESULTS AND DISCUSSION

The proposed method was evaluated on the benchmark dataset used by Benabderrahmane et al. [7], and its Pfam similarity and correlation were compared against existing techniques. The UniProt dataset [8] was used to construct positive and negative interaction sets, and Pfam similarity was calculated as the ratio of protein families shared between proteins to the total families they belong to [9]. Similarity scores are reported under Best Match

Average (BMA) Nunivers [10]. Pearson's correlation coefficient assessed the correlation between sequence similarity, from the BLAST log bit score [11], and semantic similarity.

4.1 Intraset Similarity and Semantic Similarity Evaluation

Evaluation was performed on a DAG rooted at GO:0003674, using an MF ontology snapshot with 27 direct descendants. Table 1 reports similarity for the GO term pairs GO:0046572-GO:0016829 and GO:0060089-GO:0004872, benchmarked against Resnik, Lin, Wang, and GOGO [12]. The proposed method is competitive on both pairs, though Wang and Lin attain marginally higher scores here; its advantage becomes more pronounced in the aggregate Pfam and correlation analyses that follow.

Table 1: Similarity comparison with existing methods (Resnik, Lin, Wang, GOGO) for two GO term pairs.

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

Intraset similarity was further calculated for Pfam clans (Table 2), where genes in the same clan share molecular function under the MF ontology. Figure 1 compares the intraset Pfam similarity of the proposed method with Al-Mubaid, Wang, and TopoICSim [13] across four Pfam clans; the proposed method achieves the highest similarity in every clan.

Table 2: Pfam clans used for the intraset similarity calculation.

Pfam Clan	No. of Genes
ALDH-like	18
BIR-like	9
FBD	6
Flavoprotein	7
6PGD_C	8

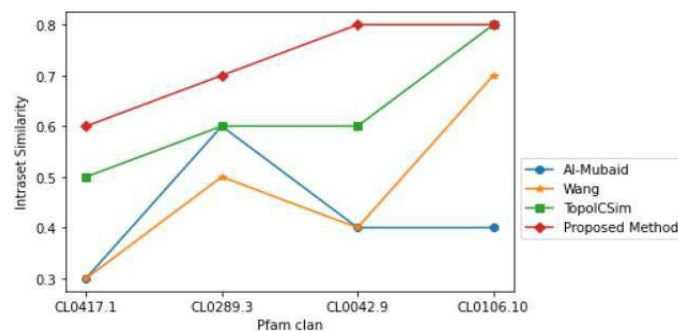


Figure 1: Pfam intraset similarity (MF ontology) by clan; the proposed method is highest in every clan.

Table 3 reports the Pearson correlation coefficient for the CC, BP, and MF ontologies. The proposed method outperforms Lord, Al-Mubaid, Wang, and TopoICSim across all three, reflecting the combined contribution of fuzzy clustering, BMA, and term depth.

Table 3: Pearson correlation coefficient for the CC, BP, and MF ontologies (highest per column in bold).

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 Method	0.654	0.623	0.714

The interaction dataset distinguishes electronically inferred annotations (IEA+) from Swiss-Prot reviewed (IEA-) annotations; combining both gives optimum coverage. Reviewed UniProt interactions (3,500) formed the

positive dataset, and unreviewed annotations formed the negative dataset. Table 4 reports the Pearson correlation between sequence and semantic similarity for MF, CC, and BP under IEA- and IEA+. The proposed method achieves the highest correlation overall.

Table 4: Pearson correlation between sequence and semantic similarity for MF, CC, and BP under IEA- and IEA+ (highest per column in bold).

Approach	MF (IEA-)	CC (IEA-)	BP (IEA-)	MF (IEA+)	CC (IEA+)	BP (IEA+)
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

CONCLUSION

This paper presents an improved semantic similarity measure for Gene Ontology terms that combines information content with topological factors of the DAG - structural depth, membership, weight, and shortest path to the MICA. The method was evaluated on benchmark datasets and Pfam clans using intraset similarity, outperforming Lord, Wang, Al-Mubaid, and TopoICSim on most comparisons, and achieved the highest Pearson correlation between sequence and semantic similarity for the Molecular Function ontology on both positive and negative datasets.

REFERENCES

- Alvarez, M. A., Qi, X., & Yan, C. (2011). A shortest-path graph kernel for estimating gene product semantic similarity. *Journal of Biomedical Semantics*, 2, Article 3. <https://doi.org/10.1186/2041-1480-2-3>
- Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L.-S. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(suppl_1), D154–D159. <https://doi.org/10.1093/nar/gki070>
- Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., & Devignes, M.-D. (2010). IntelliGO: A new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11, Article 588. <https://doi.org/10.1186/1471-2105-11-588>
- Ehsani, R., & Drabløs, F. (2016). TopoICSim: A new semantic similarity measure based on gene ontology. *BMC Bioinformatics*, 17, Article 296. <https://doi.org/10.1186/s12859-016-1160-0>
- Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1), D279–D285. <https://doi.org/10.1093/nar/gkv1344>
- Ikram, N., Qadir, M. A., & Afzal, M. T. (2018). Investigating correlation between protein sequence similarity and semantic similarity using Gene Ontology annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3), 905–912. <https://doi.org/10.1109/TCBB.2017.2695542>
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 296–304).
- Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. *Bioinformatics*, 19(10), 1275–1283. <https://doi.org/10.1093/bioinformatics/btg153>
- Mazandu, G. K., & Mulder, N. J. (2014). Information content-based Gene Ontology functional similarity measures: Which one to use for a given biological data type? *PLOS ONE*, 9(12), e113859. <https://doi.org/10.1371/journal.pone.0113859>
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv*. <https://arxiv.org/abs/cmp-lg/9511007>
- Shen, Y., Zhang, S., Wong, H.-S., & Zhang, L. (2014). Characterisation of semantic similarity on gene ontology based on a shortest path approach. *International Journal of Data Mining and Bioinformatics*, 10(1), 33–48. <https://doi.org/10.1504/IJDMB.2014.062887>
- Wang, J., Zhou, X., Zhu, J., Zhou, C., & Guo, Z. (2010). Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, 11, Article 290. <https://doi.org/10.1186/1471-2105-11-290>
- Zhao, C., & Wang, Z. (2018). GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific Reports*, 8, Article 15107. <https://doi.org/10.1038/s41598-018-33219-y>