A Review oftechniques for finding Semantic Similarity between interacting proteins

[1]GayatriC. Bagul & [2]Prof. Dr.S. M. Kamalapur
[1]M.E. Student,[2]Professor
[1]Department of Computer Engineering,
[1]K. K. Wagh Institute of Engineering Education & Research, Nashik,
SavitribaiPhule Pune University, Maharashtra, India.

Received: January 19, 2019 Accepted: March 02, 2019

ABSTRACT: *Gene Ontology is gaining attention nowadays. Gene ontology has been widely used for protein-protein interaction (PPI) networks to discover functions of genes from all the organisms. GO maintains a dynamic, structured, precisely defined and controlled vocabulary of terms to represent the roles and cellular localizations of proteins. Gene ontology terms are annotated to proteins. Given GO and its annotation data, functional closeness between two proteins is estimated by measuring their semantic similarity. The semantic similarity between two interacting proteins can be estimated by combining the similarity scores of the GO terms associated with proteins. There are several techniques available to find semantic similarity between two interacting proteins. These methods for computing the semantic similarity between GO terms are based on either information content of GO terms or topological properties of the gene ontology graph.*

Key Words: *Terms-Gene Ontology, Gene Ontology graph, information content, protein-protein interaction, semantic similarity, topological properties.*

I. Introduction

In Bioinformatics, proteins play major structural and functional roles in a cell. They are involved in all cell functions that make organs and the body work of any living things. Each protein defines some functions. Some proteins provide structure and support for cells, while others are involved in transmitting signals to coordinate biological processes between cells, tissues and organs, or in defending the body from foreign invaders such as viruses and bacteria, or in carrying out thousands of chemical reactions such as reading the genetic information stored in DNA to form new molecules.

Protein-protein interaction plays vital role in predicting the protein function of target protein [1]. This protein function needs to be defined in structured manner.Bioinformatics has increasingly become a data-intensive discipline therefore, ontologies have emerged as an essential computational tool to assist in the organization and analysis of data. Ontologies are computational structures that describe the entities and relationships of a domain of interest in a structured computable format, which allows for their use in multiple applications [2].

Currently, one of the most essential ontologies within the bioinformatics community is the Gene Ontology (GO). Gene Ontology (GO) maintains a dynamic, structured, precisely defined and controlled vocabulary of terms to represent the roles and cellular localizations of proteins. Gene Ontology (GO) is a repository of annotations of genes. The Gene Ontology provides a standard vocabulary of terms for describing gene product characteristics and gene product annotation data. Gene Ontology is organized as three independent directed acyclic graphs (DAG) based on the three aspects of proteins which advances in functional attributes (terms): Molecular Function (MF), Biological Process (BP), Cellular Component (CC). The most essential relations for GO are is_a and Part_of.

II. RELATED WORK

In this literature work, various strategies and techniques for finding semantic similarity between interacting proteins of the gene ontology graph are discussed. Gene ontology graph is utilized for finding semantic similarity between interacting proteins then topological properties are extracted from the graph and using this extracted topological properties graph is visualized.

The graph structure of the Gene Ontology (GO) allows the comparison of GO terms by semantic similarity. GO terms have been widely used for annotation purpose in different organisms based on various evidence sources. GO terms are annotated to proteins. Annotations are of two types: Manual, Electronic. These GO terms can be used to infer the functional relationships between two proteins.

Functional relationship between two proteins can be investigated by measuring their semantic similarity. The semantic similarity between interacting proteins can be computed by combining the similarity scores of GO terms that are annotated to proteins. Measuring similarity or distance between GO terms is a key step for determining hidden relationship between genes. The notion of semantic similarity between GO terms is an essential step in knowledge discovery related tasks.

A. Ontology Structure Based Approach [2]

The ontology structure-based approaches utilize the graph structure of an ontology. The ontology structure-based approaches are categorized as: edge-based approach and node-based approach.

- Edge-Based Approach

The edge-based approaches are dependent on the paths between two GO terms in the GO graph. The edge-based methods mainly measure the shortest path length between two terms. The edge-based approach calculates the shortest distance by the number of edges, between the nodes associated with two terms of the gene ontology. The shorter this distance, the more similar they are.

Edge-based approaches are based mainly on counting the number of edges in the graph path between two terms[3]. The most common technique, distance, selects either the shortest path or the average of all paths between GO terms, when more than one path exists. This edge-based technique yields a measure of the distance between two terms, which can be easily converted into a similarity measure. The common path technique calculates the similarity directly by the length of the path from the lowest common ancestor of the two terms to the root node of the Gene Ontology[4].

- Node-Based Approach

The node-based methods mainly count the number of common ancestors between two GO terms. Node-based approaches rely on comparing the properties of GO terms involved, which can be related to the GO terms themselves, their ascendants (ancestors), or their descendants. One concept commonly used in these approaches is information content (IC).

Information content gives information about GO term. Alternatively, the IC can also be calculated by the total number of children a single GO term has in the GO graph structure [5], although this approach is less commonly used. The concept of IC can be applied to those GO terms that have common ancestors to quantify the information they share and thus measure their semantic similarity between them. There are two main approaches for calculating IC are: Most Informative Common Ancestor (MICA), Disjoint Common Ancestors (DCA).

- Most Informative Common Ancestor (MICA)

In this technique only, the common ancestor with the highest IC is considered [6].

- Disjoint Common Ancestors (DCA)

In this technique all disjoint common ancestors that do not subsume any other common ancestor are considered [7].

In this ontology structure-based technique only the graph structure is utilized therefore, it is difficult to measure how specific and informative a GO term is.

B. Annotation Based Approach [3]

The annotation-based approaches use not only the graph structure of an ontology but also annotations to the terms. The specificity of a GO term in this category is usually represented as the information content of GO term. The annotation-based approaches are broadly classified in two categories: Pairwise Annotation Approach, Groupwise Annotation Approach.

- Pairwise Annotation-Based Approach

Pairwise approaches measure functional similarity between two GO terms by combining the semantic similarities between their terms. Each gene product is represented by its set of annotations, and semantic similarity is calculated between GO terms from given set.

The most common methods of measuring protein functional or semantic similarity have been pairwise approaches based on node-based technique, namely, Resnik's in 1999, Lin's, and Jiang and Conrath's. Lord et al. were the first to apply these measures, using the average of all pairwise similarities as the strategy [9].

- Groupwise Annotation-Based Approach

Groupwise approaches do not rely on combining similarities between individual GO terms. It directly computes semantic similarity by utilizing graph structure. In graph approaches GO terms are represented as the subgraphs of GO corresponding to all their respective annotations.The first graph-based similarity method to be applied to GO was Lee et al.[8]. This method uses Groupwise approach to find similarity between gene product by the number of terms they share with other GO terms present in gene ontology.

Groupwise approaches calculate GO term similarity directly by one of three approaches:Set, Graph, Vector.

- Set

Set approaches consider only direct annotations which are electronic annotations.

- Graph

In this approach gene ontology is represented as the directed acyclic graphs corresponding to all their annotations which are direct and inherited.

- Vector

In vector approaches GO terms are represented in vector space, with each term corresponding to a specific dimension, and functional similarity is calculated using vector similarity methods.
In this annotation-based technique only the Information Content (IC) of specific GO term is calculated and utilized. Gene ontology graph structure is not utilized; hence the performance is affected.
Hybrid Approach [12]
Compared with Ontology Structure based and Annotation based methods for computing the semantic similarity between GO terms are based on either the information content of GO terms or topological properties of the gene ontology graph. Since, similarity methods has been widely used in prediction purposes, such as protein-protein interaction prediction, interaction network prediction. We must ensure that these predictions are reliable and precise.

Existing methods do not provide precise and reliable protein predictions so to overcome this problem hybrid approach is the solution. Hybrid approach utilizes both the topological properties of the Gene Ontology graph and information contents of GO terms.

III. CONCLUSION

In this paper we have discussed various semantic similarity between interacting proteinstechniquesadapted in bioinformatics. These techniques have high influence on system performance. Semantic similarity finding methodshave been widely used in prediction purposes such as: Protein-protein interaction prediction, network prediction, pathway modelling etc. Existing methods are either based on topological properties of graph or information content of GO terms therefore, are not intrinsic to GO. To overcome this problem hybrid approach is the solution. Hybrid approach utilizes both the topological properties of the Go graph and information content of GO terms.Performance measures such as precision, f1-measure etc. are considered.

IV. Acknowledgment

I would like to express my gratitude to my guide Prof Dr. S. M. KamalapurProfessor, Computer Engineering, K.K.W.I.E.E.R., Nashik for giving me a moral support, valuable guidance and encouragement in making this literature survey. A special thanks to Prof.Dr. K. N. Nandurkar, Principal and Prof.Dr. S. S. Sane, Head of Department of K.K.W.I.E.E.R.,Nashik for their kind support and suggestions. It would not have been possible without the kind support. I would like to extend my sincere thanks to all the faculty members of the department of computer for their help.

References

1. V. Srinivasa Rao, K. Srinivas, G. N. Sujini, and G. N. Sunand Kumar (2014) Protein-Protein Interaction Detection: Methods and Analysis Volume 2014, Article ID 147648, 12 pages.
2. Stevens R, Goble CA, Bechhofer S (2000) Ontology-based knowledge representation for bioinformatics. Brief Bioinformatics 1(4):398–414.
3. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. pp. 17–30.
4. Wu Z, Palmer MS (1994) Verb semantics and lexical selection. Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL 1994). pp. 133–138. URL http://dblp.uni-trier.de/db/conf/acl/acl94.htmluP94.
5. Seco N, Veale T, Hayes J (2004) An intrinsic information content metric for semantic similarity in wordnet. ECAI. pp. 1089–1090.
6. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. Proc. of the 14th International Joint Conference on Artificial Intelligence. pp. 448–453.
7. Couto FM, Silva MJ, Coutinho PM (2005) Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. Proc. of the ACM Conference in Information and Knowledge Management as a short paper.
8. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085–1094.

9. Lord P, Stevens R, Brass A, Goble C (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275–1283.
10. GO-Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32: D258–D261.
11. Joslyn C, Mniszewski S, Fulmer A, Heaton G (2004) The gene ontology categorizer. Bioinformatics 20: i169–177.
12. Dutta P, Basu S, Kundu M. IEEE/ACM Trans Comput Biol Bioinform. 2018 May-Jun;15(3):839-849.doi: 10.1109/TCBB.2017.2689762. Epub 2017 Mar 31. Assessment of Semantic Similarity between Proteins Using Information Content and Topological Properties of the Gene Ontology Graph