

Gene Ontology: application and importance in functional annotation of the genomic data

Reshu Saxena^{1,2}, Ritika Bishnoi³ and Deepak Singla^{3,*}

¹Fels Institute for Cancer Research & Molecular Biology, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, United States,

²Current address: Department of Biological Sciences and Biotechnology, Institute of Advanced Research (IAR), Gandhinagar, India, ³School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India

*School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India

9.1 Background

Functional characterization of gene products is fundamental to understand the underlying biology, research advancement in areas of drug development, gene set enrichment analysis, and the biological process (BP) and other mechanistic aspects (Huang, Sherman, & Lempicki, 2009; Mi et al., 2013; Xuan et al., 2019; Zeng, Zhang, & Zou, 2016; Zhang et al., 2019). The functional characterization of a gene product in an experimental laboratory is a daunting task that needs few to several years for a gene. Thus one can assume the time required to annotate a whole genome or proteome of a species. Furthermore, manual curation of gene/protein and assigning its function based on the data mining process required the expert and dedicated biological data curator that is again a very time-consuming process. To overcome these hurdles, nowadays biologists are utilizing computational tools for accurate functional annotations of gene products that allow the prediction and characterization of the identified proteins involved in associated pathways (Jiang et al., 2016; Zhao et al., 2020). Bioinformatics tools are thus emerging as a promising approach for the functional elucidation of genes in a scientific study.

Continuous advancement in sequencing technologies and significant drop in sequencing cost are regularly generating draft/chromosome level assembly of novel organisms from simpler prokaryotes to complex eukaryotes (Levy & Boone, 2019; Levy & Myers, 2016; Straiton et al., 2019). At NCBI, more than 56,000 genome assemblies are present in complete or draft level of which ~14,000 are from eukaryotes, ~282,000 are from prokaryotes, and 41,000 are from viruses (<https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/>). Manual or experimental annotation of such huge data is not possible, further the amount of time and computational skills required for large eukaryotic genomes pose another challenge to the genome annotation process. In the absence of any comprehensive source of standard biological databases, the rapid development of new instruments and advancements in techniques further amplify these challenges. In this situation, accurate computational tools and databases represent the best alternative to speed up the process of structural and functional annotation.

Functional genomics research has widened significantly utilizing Gene Ontology (GO) as a comprehensive base (Ashburner et al., 2000; Dessimoz & Škunca, 2017; Zhao et al., 2020). GO terminology is commonly used, most comprehensive, and widely accepted as standards describing functional attributes of gene products. It is a graph-based technique that classifies the gene product in a hierarchical based graph structure. The main objective of creating the GO-based annotation is to develop a unified biological nomenclature system that can best describe the structural and functional characteristics of a gene or gene product from any organism (Balakrishnan et al., 2013). GO database represents the world's largest information in both human and machine readable form that could be used for large-scale computational analysis in genomics projects (Carbon et al., 2019). GO is being used for the functional annotation, by arranging GO terms in a hierarchy and representing their connections. GO database periodically updates the GO terms through

the GO consortium. With large input datasets and their parental relationships depiction, the directed acyclic graph structures of GO give an insightful outline of the gene annotation (Ashburner et al., 2000).

GO is intended to summarize the interrelatedness of input genes and the biological terms for the analysis of genomic data in the functional aspect of the biological study. GO resources include comprehensive data that comprise known information about the genes and gene products covering molecular aspects across all life forms. The recent release of database hold 45,000 terms [29,698 for BPs; 11,147 for molecular functions (MFs); and 4201 for cellular components (CCs)] that are linked with approximately 134,000 relationship. This database could be used to predict the gene function through a computational approach in a systematic manner by the analysis of functional attributes of the genome across species. The tools for searching the GO database assign a particular GO term that is supported by any of the following levels of evidence, such as experimental evidence, based on computational analysis, curation, author statement, genomic evidence, or combinatorial evidence. GO annotation uses a standard tab-separated file format called Gene Associations File (GAF, <http://www.geneontology.org/GO.format.annotation.shtml>) to record the information systematically. GAF format has some mandatory fields, such as DB, GO ID, and evidence code, and some are optional fields, such as annotation extension, gene product id, DB object name, and synonymous. Besides this, some specific databases have been constructed in the past, such as agriGO and Plant Ontology (PO) (Cooper & Jaiswal, 2016; Tian et al., 2017). The PO database uses the controlled, structured vocabulary to describe plant anatomy, morphology, and the stages of plant development that could be exploited for comparative genomics. Initially, the database was started with functional annotation for rice, maize, and *Arabidopsis* but now covers a wide range of plants from green algae to angiosperms. Similarly, agriGO has been developed to support the GO analysis on plants and agricultural species. AgriGO database holds the information on ~400 species that are further classified into crops, vegetables, fish, birds, and insects that highlights the importance and use of this database.

9.2 Gene Ontology–based classification

Ontology is a representation of information that we currently have for a given domain. Ontologies consist of a set of classes and the relationships operating within these classes (Hill et al., 2008). A class represents the basic unit within ontology, for example, cancer, apoptosis, and photosynthesis. Each class is further associated with textual information called Metadata that might be associated with other secondary identifiers. Metadata includes cross-references with other database and resources that further provide the evidence of association. The classes are linked with each other in the form of a relationship with the help of acyclic graphs. For example, class “cancer” is associated with apoptosis, cell death, etc. The classes or terms in GO have a name, a unique alphanumeric identifier, a known description, and a function with a domain linked. These representations of terms are not universal but vary across the species and depend on the concerned research area. The GO project describe properties of gene products defined for a biological domain. GO is a major bioinformatics resource that merges gene and gene products with a focus on their functions across the species. Each term represents the defined relationships to other terms within and across the domains too. The GO elucidates the knowledge of the biological domain, covering basically three aspects (Fig. 9.1): (1) CCs; (2) MFs; and

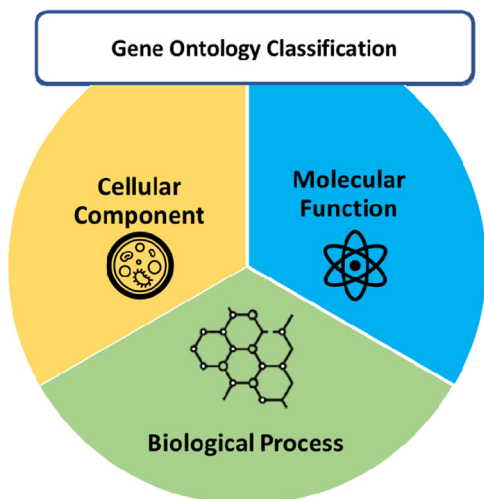


FIGURE 9.1 Representation of the Gene Ontology term classification system.

(3) BPs. Each of these three components encapsulated different biological information that annotates a gene or gene product.

9.2.1 Cellular component

CC is useful in identifying the subcellular location of gene/protein, such as cytoplasmic, membrane, mitochondrial, chloroplast, or nuclear. The CC represents the cell constituents or the extracellular environment. This class in GO term refers to the cellular anatomy and not the processes, that is, the cellular location or the structure involved in a particular function performed by a gene product. Most of the proteins are synthesized by the nuclear genome but many of them show their presence in other subcellular locations due to the presence of import/export signal. This is of particular importance when the aim is to delineate the location of proteins at the genome level.

9.2.2 Molecular function

MF represents the elemental molecular level activities performed by gene products. Molecular components are helpful in predicting the MF of gene products, such as involvement in enzymatic activity, transcription factor binding, or DNA/RNA binding. GO terms give knowledge of the functions or activities performed, but not the entities performing that action. It can be broadly divided into two parts: (1) biochemical activities, and (2) function as a component of a large system or process. These MFs are attributed to either individual gene products or sometimes to molecular complexes composed of several gene products. For example, the term “biochemical activity” includes enzymatic activity, nucleotide binding, etc., and similarly the term “function as receptor” means that the term is the component of a large complex or process.

9.2.3 Biological process

Likewise, the third component, that is, BP provides information regarding the biological or metabolic pathways in which a particular gene or protein is involved. BPs represent the network of molecular events related to the functional aspects of living organisms and their units, for example, protein phosphorylation, WNT signaling, and hormone signaling. Multiple molecular operations are involved in a concerted BP, such as signal transduction or DNA repair. GO only indicates a broad BP but does not determine a complete pathway along with underlying dynamics and mechanisms. In the acyclic graph, each of these terms is further subdivided into much smaller and informative components.

9.3 Annotation of unknown gene/genome

The continuous drop in the cost of sequencing and highly efficient sequencing technologies has allowed scientists to sequence and characterize the novel genome. Several sequencing and resequencing genome projects are in progress and have generated a huge volume of digital data for in-depth analysis. The most critical part is the structural and functional annotation that can be made possible with computational analysis. To annotate a gene/genome more precisely, GO annotation played an important role. The manual annotation based on scientific evidence is the best practice to accurately annotate a gene product. But first, an expert curator requires a lot of effort and time in collecting and reviewing the literature, followed by an annotation. Second, it is not possible to get literature for every single gene; thus a significant portion remains unannotated. The second approach is a computational prediction based on similarity search. Although the computational method itself depends upon the experimental evidence with at least one similar protein experimentally annotated previously, still it has been widely used as the best alternative strategy. BLAST/PSI-BLAST and Hidden Markov Model (HMM) are most commonly used for structural and functional assignment. This is the simplest method used to transfer the annotation from experimentally validated hits based on *e*-value and query coverage (Fig. 9.2). However, this sometimes leads to erroneous results due to mutation in key residues or changes in domain architecture. To overcome this issue, recently, deep-learning- or machine-learning-based tools have been implemented, which are much fast and reliable. Here, we briefly review and highlight the application of some commonly used software's/web servers developed in the past for GO-term annotation of unknown sequences (Table 9.1).

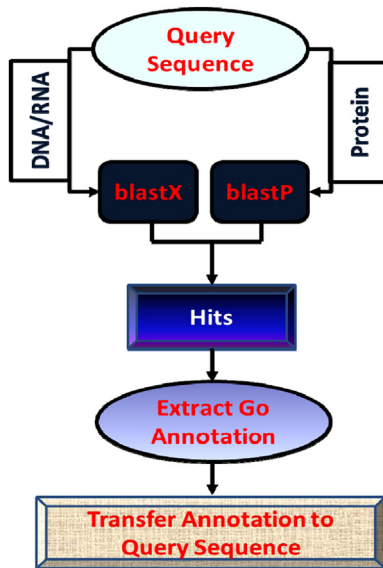


FIGURE 9.2 Flow diagram for the transfer of Gene Ontology term to an unknown gene.

TABLE 9.1 Description of tools/software's used for GO-term annotation.

Software/tools	URL	References
Blast2GO	https://www.blast2go.com	Conesa and Götz (2008)
IPRscan	http://www.ebi.ac.uk/interpro/search/sequence	Jones et al. (2014)
GAAP	http://www.deepaklab.com/gaap	Unpublished
GOnet	http://tools.dice-database.org/GOnet/	Pomaznoy, Ha, and Peters (2018)
DEEPred	https://github.com/cansyl/DEEPred	Sureyya Rifaioglu et al. (2019)
SDN2GO	https://github.com/Charrick/SDN2GO	Cai, Wang, and Deng (2020)
AmiGO	http://amigo.geneontology.org/amigo	Carbon et al. (2009)
OBO-edit	http://oboedit.org	Day-Richter et al. (2007)
GO-FEAT	http://computationalbiology.ufpa.br/gofeat	Araujo et al. (2018)

Note: GAAP, Genome Assembly and Annotation Package; GO, Gene Ontology; IPRscan, InterProScan.

9.3.1 Blast2GO

Blast2GO is a comprehensive commercial software built on the JAVA platform for the functional annotation of genes/genome. It is a very versatile, user-friendly bioinformatics tool with wide-ranging functions including GO annotation and gene set enrichment analysis (GSEA). It uses an algorithm based on similarity, homology, and GO hierarchy to optimize functional attributes derived from homologous sequence data. It supports several other software for improved statistical analysis and management of the annotation results.

9.3.2 IPRscan (InterProScan)

It is a JAVA-based free software based on HMM profiles and widely used for domain/motif-based protein function prediction. But, along with this, the software will extract the GO term from identified hits and could be used for assigning the GO terms. Developments include improvements and additions to the outputs of the software and the complete reimplementation of the software framework, resulting in a flexible and stable system that can use both multiprocessor machines and/or conventional clusters to achieve scalable distributed data analysis.

9.3.3 Genome Assembly and Annotation Package

This is platform-independent software designed specifically for the de novo genome assembly and annotation purpose. It supports quality control, read quality analysis, and de novo assembly. Along with this, users can go for prokaryotic gene prediction and genome and proteome annotation with Blast and InterProScan, GO term, Enzyme, and pathway analysis. To assign the GO term to the predicted gene/protein, the software extracts the GO term from InterProScan hits and assigns it to the query sequence.

9.3.4 GOnet

This software can be used to visualize the interaction and relationships between terms and genes. It allows the user to perform GO term annotation analysis, produces an interactive display of the results, and provides information on the functional aspects of the genes, such as how they are interconnected, which allows better biological interpretation.

9.3.5 DEEPred

It is a very promising and multitasking tool for accurate protein function prediction and overcomes the limitations of conventional algorithms for the same. It is optimized by considering the hierarchical stack of GO terms and uses hyper-parameter tests to assess the larger data. DeePred is specifically useful when the data set is huge and noisy as it uses deep-learning-based neural network algorithms that provide a significantly accurate prediction of protein function.

9.3.6 SDNGO

This is another deep-learning-based classification model for protein function prediction using protein sequences, domains, and networks. It further integrates this information to accurately predict protein function depicted by GO terms.

9.3.7 AmiGO

This is a web-based tool for searching and browsing the GO database and allows visualizing ontologies and annotation of gene products. It enquires the GO terms directly for the analysis of huge gene sets data by using the BLAST tool. The information contained in the GO consortium can be directly accessed online by AmiGO at the GO website or it can be used offline by downloading and installation being an open-source software.

9.3.8 OBO-Edit

OBO-Edit is open-source software used as an ontology editor implemented in JAVA. This is developed by Berkeley Bioinformatics and maintained by the GO consortium. It is easy and simple to use with potential searching capacity and employs a graph-based approach for editing and data display. This user-friendly editing tool is freely available and also customizable according to user requirements. It also features a reasoner that provides more concrete information considering the relationships and properties otherwise neglected.

9.3.9 Gene Ontology Functional Enrichment Annotation Tool

GO-FEAT (Gene Ontology Functional Enrichment Annotation Tool) is a freely available, user-friendly, easily customizable web server and emerging as an integrated tool to analyze functional annotations and enrichment of transcriptome sequence data. It overcomes the limitations of several other software that are either not free or have limited sequence input capacity and lack proper visualization and management. It makes downstream analysis easy by providing multiple approaches to export results in the form of graphs, charts, reports, or tables depending on the need of the user.

9.4 GO enrichment analysis

High-throughput approaches generate a huge data yield set of genes and therefore need methods to retrieve significantly enriched functional attributes of these genes in the context of the disease process, specific conditions/treatment, etc. (Costa-Silva, Domingues, & Lopes, 2017; Nagalakshmi et al., 2008). It gets further complicated to interpret such biological data due to more data noise and limited approaches to understand the underlying biology. To better decipher

such large datasets, a comparative analysis of the input gene set with terms in GO is required followed by a statistical test to see how enriched the input genes are in each term. Like, transcriptome data are routinely used for identifying the differentially expressed transcripts between different conditions and identified transcripts need to be organized and prioritized based on the significantly associated GO term (Lee et al., 2014; Yang et al., 2016, 2018). To characterize the disease phenotypes and the associated genes and proteins, GSEA is employed (Subramanian et al., 2005). Using statistical strategies differentially expressed group of genes is identified from a large sets of genes and significantly enriched GO term is identified based on a statistical test that links the particular gene or set of genes responsible for a disease condition (Li et al., 2011; Osborne et al., 2009; Yu et al., 2015).

“Overrepresentation Analysis” (ORA) approach was developed to address this that incorporates large lists of genes from experiments and determine gene sets with significant enrichment sharing a common theme (Boyle et al., 2004). This indicates the pathways associated with the input genes. ORA being a simple and easy approach is widely used but has certain limitations. First, it does not consider gene networks and their influence on each other. Second, it also has risks of excluding important genes that are unable to reach the required statistical cut off (Khatri & Drăghici, 2005; Khatri, Sirota, & Butte, 2012). This turns to be unrealistic for understanding the scientific question under study and may lead to misinterpretation. GSEA is an advanced second-generation pathway analysis tool, also mentioned as a significant analysis of function and expression (Barry, Nobel, & Wright, 2005). This approach overcomes the limitations by making use of complete gene sets including the gene expression quantitatively. This method maps the list of genes from datasets against the pathways and makes a comprehensive network with relevance to the study. These calculate the statistical significance at both gene level and pathway level and determine it through multiple testing (Ackermann & Strimmer, 2009).

Enrichment analysis results in the identification of over/underrepresented GO terms that ultimately connect these with genes and establish the relationship of genes with specific conditions or phenotype (Tokar et al., 2020). In the background, the software/web server has a precalculated GO term for each gene. When the user provides their list to the software, it will calculate the GO term for each gene and then complete statistical analysis based on number of genes associated with a particular GO term in the user list and number of genes present in the database for that particular GO term will be done. Finally, the results are shown in a tabular format indicating whether the GO term is over- or underrepresented along with statistical *P*-value. Gotcha is one of the first tools that assigns weight to each GO term, calculates the statistical score, and establishes semantic relationships (Martin, Berriman, & Barton, 2004). Another tool PFP also follows a similar approach but it uses PSI-BLAST instead of BLAST to find distant homologs (Hawkins et al., 2009). In this section, we briefly review the software/tools routinely used for GO-term enrichment analysis. These software/tools are also listed in Table 9.2 with their respective URLs.

9.4.1 DAVID (Database for Annotation, Visualization, and Integrated Discovery)

DAVID, developed by the Laboratory of Immunopathogenesis and Bioinformatics, NIAID, is a freely available, most commonly used, online tool for gene functional classification. It is a reliable source to provide the fast organization of gene lists into groups with functional relatedness. This allows uncovering the biological relevance of the study results. This allows the users to provide a list of genes as input and then select any of the annotation categories from various sources and finally provide the list of significantly enriched GO terminology.

9.4.2 PANTHER (Protein ANalysis THrough Evolutionary Relationships)

This software is the part of Gene Ontology Phylogenetic Annotation Project and is designed to classify proteins (and their genes) to facilitate high-throughput analysis. The classification criteria depend on families considering the evolutionarily related proteins, functional similarity, MF, BPs connecting the genes interacting for the same process, and further the pathways that specify the connection of interacting molecules. Protein ANalysis THrough Evolutionary Relationships uses a comprehensive protein library combined with human-curated pathways and evolutionary ontology. If a gene is not in the library, it is classified based on its protein sequence conservation and by finding a related gene.

9.4.3 g:Profiler

g:GOSt is the part of g:Profiler that performs functional enrichment analysis, also known as overrepresentation analysis (ORA) or gene set enrichment analysis based on the input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms. The software regularly updates the data from the Ensembl database and parasite-specific data from WormBase. In addition to GO, information is taken from KEGG,

TABLE 9.2 List of software for GO-term enrichment analysis.

Software/tool	URL	References
DAVID	https://david.ncicrf.gov	Huang et al. (2009)
PANTHER	http://pantherdb.org	Mi et al. (2013)
g:Profiler	http://biit.cs.ut.ee/gprofiler	Raudvere et al. (2019)
ClusterProfiler	https://github.com/GuangchuangYu/clusterProfiler	Yu et al. (2015)
Enrichr	http://amp.pharm.mssm.edu/Enrichr	Chen et al. (2013)
ToppGene	https://toppgene.cchmc.org	Chen et al. (2009)
QuickGo	http://www.ebi.ac.uk/QuickGO	Binns et al. (2009)
REVIGO	http://revigo.irb.hr	Supek et al. (2011)
WEGO	http://wego.genomics.org.cn/cgi-bin/wego/index.pl	Ye et al. (2018)
ShinyGO	http://bioinformatics.sdstate.edu/go	Ge et al. (2020)
ViSEAGO	https://bioconductor.org/packages/ViSEAGO	Brionne, Juanchich, and Hennequet-Antier (2019)
PoGO	http://bioinformatica.vil.usal.es/lab_resources/pogo	Jung et al. (2010)
GORilla	http://cbl-gorilla.cs.technion.ac.il	Eden et al. (2009)
EasyGO	http://bioinformatics.cau.edu.cn/neweasygo/	Zhou and Su (2007)
GOEAST	http://omicslab.genetics.ac.cn/GOEAST/	Zheng and Wang (2008)
GOAT	http://goat.man.ac.uk/	Xu and Shaulsky (2005)
GOLEM	http://reducio.princeton.edu/GOLEM/	Sealfon et al. (2006)
GOsTo	http://www.paccanarolab.org/gosstoweb/	Caniza et al. (2014)
NaviGO	http://kiharalab.org/web/navigo	Wei et al. (2017)

Note: Bold terms highlight the software/application that could be used for GO term annotation. *DAVID*, Database for Annotation, Visualization, and Integrated Discovery; *GO*, Gene Ontology; *GOEAST*, Gene Ontology Enrichment Analysis Software Toolkit; *GOAT*, Gene Ontology Annotation Tool; *GOLEM*, Gene Ontology Local Exploration Map; *GORilla*, Gene Ontology enRichment anaLysis and visualizAtion tool; *GOsTo*, Gene Ontology semantic similarity Tool; *PANTHER*, Protein ANalysis THrough Evolutionary Relationships; *PoGO*, Prediction of Gene Ontology; *REVIGO*, Reduce & Visualize Gene Ontology; *WEGO*, Web Gene Ontology.

WikiPathways, miRTarBase, TRANSFAC, Human Protein Atlas, protein complexes from CORUM, and human disease phenotypes from Human Phenotype Ontology. The web server currently supports around 500 organisms and accepts hundreds of identifier types.

9.4.4 clusterProfiler

clusterProfiler is an R package for overrepresentation and gene set enrichment analysis for several curated gene sets. It generates a workflow combining analysis and visualization module that is extendable across species and supports analysis for humans, mice, and yeast. It also allows users to compare the results of analyses performed on several gene sets.

9.4.5 Enrichr

It is an integrative open-source JAVA-based web application for enrichment analysis that ranks enriched terms in the form of new gene sets and represents the enrichment results utilizing several interactive visualization approaches. It is open-source and freely available tool.

9.4.6 ToppGene

It is a freely available online tool for gene list enrichment analysis of different categories, such as GO terms, chromosomal locations, and disease associations. Based on the functional attributes and associated protein networks, the tool builds the gene hierarchy relevant to the study.

9.4.7 QuickGo

QuickGo is a web server that provides simple browsing of GO. It uses extensive computational filters that offer bulk generation of specific subsets of GO annotations by customizable sequence mapping. GO slims are used, which is a collective list of GO's full set of terms available from the GO project.

9.4.8 REVIGO (Reduce & Visualize Gene Ontology)

REViGO is a web resource developed by Rudjer Boskovic Institute, Croatia, that can summarize and visualize long lists of GO terms after removing unnecessary GO terms. It is a freely available tool that uses a simple clustering algorithm depending on semantic similarity measures and similarity-based scatterplots, interactive graphs, or tag clouds for visualization.

9.4.9 WEGO (Web Gene Ontology)

WEGO is a freely available, useful tool for plotting the GO annotation results used widely in many important biological, both plants and animals, research projects. It has become a popular tool for downstream gene annotation analysis and comparative genomics studies.

9.4.10 ShinyGO

It is a tool to improve graphical visualization and to generate an Interactive networks annotation database that inputs data to get enriched GO terms and pathways across several plant and animal species, based on annotation from Ensembl. An additional 2031 genomes, including bacteria and fungi, are annotated based on the STRING database. Also, it produces KEGG pathway diagrams with the genes highlighted, hierarchical clustering trees and networks summarizing overlapping terms/pathways, protein–protein interaction networks, gene characteristics plots, and enriched promoter motifs.

9.4.11 ViSEAGO

This tool intends to cover the three major concepts of the analysis: Visualization, Semantic similarity, and Enrichment Analysis of Gene Ontology. It is a tool for clustering biological functions using GO and semantic similarity that allows studying large-scale datasets and visualizes GO profiles for understanding the biological significance of the study. Modulating available R packages, ViSEAGO, provides functional coherence to conventional methods for functional GO analysis by accounting for closely related biological themes across multiple datasets.

9.4.12 PoGo (Prediction of Gene Ontology)

This software is based on the assignment of GO term on pattern recognition methods and specifically designed for fungal proteins. It is a meta-classifier that considers each GO class independently, classified at two base levels, and finally, one of it is used to train the meta-classifier that assigns the particular GO term.

9.4.13 GOrilla (Gene Ontology enRICHment anaLysis and visualiZAtion tool)

The software was developed at Laboratory of Computational Biology, Israel Institute of Technology. GOrilla is used for identifying and visualizing the enrichment of GO terms of the ranked gene lists. It provides good pictures by two modes either by looking for enriched GO terms appearing on the top of gene lists or searching for enriched GO terms of gene lists relative to a background gene list.

9.4.14 EasyGO

EasyGO, developed by China Agricultural University, is a user-friendly, updated, a freely available tool for gene enrichment analysis. It is intended to identify enriched GO terms from a list of microarray probe sets and also designed to provide a GO annotation database. It is specific for agronomical species, supporting 30 species.

9.4.15 GOEAST (Gene Ontology Enrichment Analysis Software Toolkit)

GOEAST, developed by the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, is a web resource, easy to use, supporting easy visualization, and analysis of the comprehensive and unbiased GO for a large volume of genomic data, such as from microarray hybridization experiments. It makes use of statistical tests to identify significantly enriched GO terms from the given gene sets.

9.4.16 GOAT (Gene Ontology Annotation Tool)

GOAT, developed by the University of Manchester is a freely available project that aims to provide biomedical researchers with a comprehensive tool for gene annotation with the GO terms.

9.4.17 GOLEM (Gene Ontology Local Exploration Map)

GOLEM is an interactive tool that allows easy navigation and visualization and analyzes the GO hierarchical structure and annotation. The visualization is graph-based, and the user can see and navigate through the graph around the GO term of interest along with GO enrichment and the underlying relationships among these. It is a freely available downloadable version.

9.4.18 GOsTo (Gene Ontology semantic similarity Tool)

GOsTo is a web-based tool that provides calculation of the semantic similarity between GO terms. It can also be downloaded as standalone software.

9.4.19 NaviGO

It is freely available web-based software for visualization and analysis of the functional similarity and relationships of GO terms and genes. Interactive navigation allows real-time visualization of the functional associations of the GO terms and genes providing statistically significant enriched functions

The visualization of high volume GO term data is very difficult. In the past, space-filling techniques were developed to represent the data in the form of a tree graph in which each term is represented in the form of a rectangular box and connected with their child/parent term. Although it provides a good overview but many times missed the labels due to space problem. To overcome this issue, recently, a circular graph-based software called CirGo has been developed, which provides a better representation of the data in the form of 2D graphs ([Kuznetsova et al., 2019](#)).

9.5 Applications

GO is widely accepted by researchers for the analysis of high-throughput studies, for deducing the huge data to more purposeful data, and for understanding mechanistic aspects of the study by establishing a disease–gene relationship. GO helps in the identification of novel genes, the alterations in their expression, distribution, and function under a different set of conditions, such as diseased versus healthy. The functional characterization of genes by GO involves gene set enrichment analysis that calculates modulation, that is, enrichment or reduction for the GO terms involved. The hierarchical arrangement of the GO terms indicates this alteration and further statistical analysis gives the most significant set of genes representing the change. However, it is critical for the analysis to use a reference set comprised of the genes only observed during the study and not the total genes. Also, it is crucial to cover both upregulated and downregulated gene sets to get unbiased results.

GO annotation to predict gene function involves a computational approach by grouping genes on the basis of something common they share. Comparable expression of genes or common biological pathways they are involved in indicates their association. However, these are to be verified experimentally to reduce the outliers and make proper gene interpretation. GO also serves to categorize the genes based on the relativeness of GO terms among gene sets. GO slim ontology is the approach to categorize the results on functionality by mapping the genome annotations to high-level GO terms. This mapping indicates the differential expression, distribution, or functionality of the gene of interest.

In 2009 Osborne et al. used Unified Medical Language System to find a disease–gene relationship based on the GO terms ([Osborne et al., 2009](#)). A GO subset called Disease Ontology has been developed that has information about the gene associated with diseases based on the GO term. In 2019 a unified resource DisGeNET database has been

developed that contained 1,134,942 gene–disease associations representing 21,671 disease-related genes (Piñero et al., 2020). Similarly, Rat Disease Ontology has been implemented in the Rat Genome database (RGD) that used a controlled vocabulary and established the gene–disease relationship (Hayman et al., 2016). Presently RGD has the 10 most comprehensive disease–gene related datasets that could be helpful for in-depth analysis. Likewise, Plant Stress Ontology, Plant–Pathogen Ontology, Plant-Associated Microbe Gene Ontology, and PO have been developed that could be used for agricultural research (Cooper & Jaiswal, 2016; Torto-Alalibo, Collmer, & Gwinn-Giglio, 2009). These kinds of databases are extremely valuable for the fast and accurate identification of disease-related genes from high-throughput sequencing at the population level (Davis et al., 2016; LePendou, Musen, & Shah, 2011).

9.6 Future prospects

Presently GO is very popularly used in genome/transcriptome annotation, functional assignment, subcellular location prediction, etc. To reduce the need for wet-lab experiments, GO tools should have more evidence for gene functions and give a more accurate and high-quality functional prediction. In the future, it will become the first choice of researchers for immediate annotation of the novel genome because of its simplicity. Although every effort has been done to keep the data simple and normalized, still the data are biased and heterogeneous in many aspects (Gaudet & Dessimoz, 2017). One such source of biasness is that all part of GO terms is not annotated at the same level. Another source of biasness is the evidence codes that are informative but cannot be used to ignore poor confidence evidence. Differences in the frequency of GO term between different species could be attributed, due to their different natures, such as mice for toxicological studies, zebrafish for developmental biology, and plants for photosynthesis. Thus it is difficult to avoid the problem that arises in statistical evaluation or interpretation during enrichment analysis. Besides that, another source of biasness is author or annotator biasness. Also, the ontology itself is not complete, it is continuously evolving and new terms are regularly added. Despite the large-scale data curation, still researchers feel that each gene product is not annotated. This requires a high-quality, complete, or near-complete annotation database in which at least one gene from every possible class or subclass is deeply annotated without any biasness. Such kind of data will be very helpful in the development of fast and accurate software/tools, precise assignment of function, and enrichment analysis.

Although software has been developed to assign the GO term to any unknown gene or protein and continuous efforts are being made on automatic assignment of GO terms, still there is scope for further improvement and more efforts are required to reach accuracy and reliability. Besides a regular update of the GO database, there is a need to develop some broad category databases specific for plants, animals, fungus, and bacteria. These types of specific and focused databases will allow researchers to get fast and accurate results. The second most important requirement is to generate a negative annotation dataset that could be used to identify and exclude irrelevant noisy annotation. This kind of dataset will be helpful to minimize the false-positive hits and unbiased annotation. Also, negative annotation data will help the researchers to develop highly sophisticated, fast, and more accurate machine-learning-based algorithms to classify the unknown dataset. Recently, deep-learning-based algorithms, such as DEEPred and SDN2GO, have been developed to annotate the unknown sequence based on the GO term. In the absence of any true negative dataset, DEEPred used a training dataset based on UniProt-KB and their associated GO terms. Similarly, SDN2GO has used information about protein–protein interactions, protein domain, and GO term to develop a deep-learning-based model. But in our opinion, the negative dataset used in this software is not the true representation; thus a highly accurate negative dataset is the immediate need to develop high-throughput models in the future. Also, the development of software/tools for better visualization of GO data will have a direct impact on understanding and interpretation of high-throughput data.

9.7 Conclusion

GO term can be used to easily assign the function to an unknown gene. In this chapter, we have covered various software and tools that could be used for predicting the function of a gene as well as for enrichment analysis. GO term enrichment analysis will help the researchers to delineate the genes highly enriched for a particular process or function that are specific to a dataset and will be helpful in the comparative genome or transcriptome analysis.

Acknowledgment

The authors are thankful to the Department of Biotechnology (DBT) (project BTISNET) for providing the bioinformatics facilities at the School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana.

Conflict of interest

The manuscript is seen and approved by all the authors and there is no competing interest.

Author contributions

RS, RB, and DS wrote the manuscript. RS and DS prepared the figures. RB prepared the list of software for GO analysis. DS conceived and supervised the project.

References

- Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, *10*(1), 47. Available from <https://doi.org/10.1186/1471-2105-10-47>.
- Araujo, F. A., et al. (2018). GO FEAT: A rapid web-based functional annotation tool for genomic and transcriptomic data. *Scientific Reports*, *8*(1). Available from <https://doi.org/10.1038/s41598-018-20211-9>.
- Ashburner, M., et al. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. Available from <https://doi.org/10.1038/75556>.
- Balakrishnan, R., et al. (2013). A guide to best practices for gene ontology (GO) manual annotation, Database. Oxford University Press. doi: 10.1093/database/bat054.
- Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, *21*(9), 1943–1949. Available from <https://doi.org/10.1093/bioinformatics/bti260>.
- Binns, D., et al. (2009). QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics (Oxford, England)*, *25*(22), 3045–3046. Available from <https://doi.org/10.1093/bioinformatics/btp536>.
- Boyle, E. I., et al. (2004). GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715. Available from <https://doi.org/10.1093/bioinformatics/bth456>.
- Bronne, A., Juanchich, A., & Hennequet-Antier, C. (2019). ViSEAGO: A bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. *BioData Mining*, *12*(1), 16. Available from <https://doi.org/10.1186/s13040-019-0204-1>.
- Cai, Y., Wang, J., & Deng, L. (2020). SDN2GO: An integrated deep learning model for protein function prediction. *Frontiers in Bioengineering and Biotechnology*, *8*, 391. Available from <https://doi.org/10.3389/fbioe.2020.00391>.
- Caniza, H., et al. (2014). GOssTo: A stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics*, *30*(15), 2235–2236. Available from <https://doi.org/10.1093/bioinformatics/btu144>.
- Carbon, S., et al. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics (Oxford, England)*, *25*(2), 288–289. Available from <https://doi.org/10.1093/bioinformatics/btn615>.
- Carbon, S., et al. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. Available from <https://doi.org/10.1093/nar/gky1055>.
- Chen, E. Y., et al. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 128. Available from <https://doi.org/10.1186/1471-2105-14-128>.
- Chen, J., et al. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(Suppl. 2), W305. Available from <https://doi.org/10.1093/nar/gkp427>.
- Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, *2008*, 619832. Available from <https://doi.org/10.1155/2008/619832>.
- Cooper, L., & Jaiswal, P. (2016). The plant ontology: A tool for plant genomics. *Methods in Molecular Biology*, 89–114. Available from https://doi.org/10.1007/978-1-4939-3167-5_5.
- Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One*, *12*(12), e0190152. Available from <https://doi.org/10.1371/journal.pone.0190152>, Edited by Z. Wei. Public Library of Science.
- Davis, A. P., et al. (2016). Generating gene ontology-disease inferences to explore mechanisms of human disease at the comparative toxicogenomics database. *PLoS One*, *11*(5). Available from <https://doi.org/10.1371/journal.pone.0155530>.
- Day-Richter, J., et al. (2007). OBO-Edit - An ontology editor for biologists. *Bioinformatics*, *23*(16), 2198–2200. Available from <https://doi.org/10.1093/bioinformatics/btm112>.
- The Gene Ontology handbook. In C. Dessimoz, & N. Škunca (Eds.), *Methods in molecular biology*. New York, NY: Springer New York. Available from <http://doi.org/10.1007/978-1-4939-3743-1>.
- Eden, E., et al. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. Available from <https://doi.org/10.1186/1471-2105-10-48>.
- Gaudet, P., & Dessimoz, C. (2017). *Gene ontology: Pitfalls, biases, and remedies*. *Methods in Molecular Biology* (pp. 189–205). Humana Press Inc. Available from http://doi.org/10.1007/978-1-4939-3743-1_14.
- Ge, S. X., et al. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, *36*(8), 2628–2629. Available from <https://doi.org/10.1093/bioinformatics/btz931>.

- Hawkins, T., et al. (2009). PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins: Structure, Function, and Bioinformatics*, 74(3), 566–582. Available from <https://doi.org/10.1002/prot.22172>.
- Hayman, G. T., et al. (2016). The disease portals, disease-gene annotation and the RGD disease ontology at the rat genome database, Database. Oxford University Press. doi: 10.1093/database/baw034.
- Hill, D. P., et al. (2008). Gene Ontology annotations: What they mean and where they come from. *BMC Bioinformatics*, S2. Available from <https://doi.org/10.1186/1471-2105-9-S5-S2>.
- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44–57. Available from <https://doi.org/10.1038/nprot.2008.211>.
- Jiang, Y., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*. Available from <https://doi.org/10.1186/s13059-016-1037-6>.
- Jones, P., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9), 1236–1240. Available from <https://doi.org/10.1093/bioinformatics/btu031>.
- Jung, J., et al. (2010). PoGO: Prediction of Gene Ontology terms for fungal proteins. *BMC Bioinformatics*, 11(1), 215. Available from <https://doi.org/10.1186/1471-2105-11-215>.
- Khatri, P., & Drăghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, 3587–3595. Available from <https://doi.org/10.1093/bioinformatics/bti565>.
- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*. Available from <https://doi.org/10.1371/journal.pcbi.1002375>.
- Kuznetsova, I., et al. (2019). CirGO: An alternative circular way of visualising gene ontology terms. *BMC Bioinformatics*, 20(1), 84. Available from <https://doi.org/10.1186/s12859-019-2671-2>.
- Lee, J.-H., et al. (2014). Thoroughbred horse single nucleotide polymorphism and expression database: HSDB. *Asian-Australasian Journal of Animal Sciences*, 27(9), 1236–1243. Available from <https://doi.org/10.5713/ajas.2013.13694>.
- LePendu, P., Musen, M. A., & Shah, N. H. (2011). Enabling enrichment analysis with the Human Disease Ontology. *Journal of Biomedical Informatics*, 44(Suppl. 1), S31–S38. Available from <http://doi.org/10.1016/j.jbi.2011.04.007>.
- Levy, S. E., & Boone, B. E. (2019). Next-generation sequencing strategies. *Cold Spring Harbor Perspectives in Medicine*, 9(7). Available from <http://doi.org/10.1101/cshperspect.a025791>.
- Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17(1), 95–115. Available from <http://doi.org/10.1146/annurev-genom-083115-022413>.
- Li, J., et al. (2011). DOSim: An R package for similarity between diseases based on Disease Ontology. *BMC Bioinformatics*, 12. Available from <https://doi.org/10.1186/1471-2105-12-266>.
- Martin, D. M. A., Berriman, M., & Barton, G. J. (2004). GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5. Available from <https://doi.org/10.1186/1471-2105-5-178>.
- Mi, H., et al. (2013). Large-scale gene function analysis with the panther classification system. *Nature Protocols*, 8(8), 1551–1566. Available from <http://doi.org/10.1038/nprot.2013.092>.
- Nagalakshmi, U., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881), 1344–1349. Available from <https://doi.org/10.1126/science.1158441>.
- Osborne, J. D., et al. (2009). Annotating the human genome with Disease Ontology. *BMC Genomics*. Available from <https://doi.org/10.1186/1471-2164-10-S1-S6>.
- Piñero, J., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1), D845–D855. Available from <https://doi.org/10.1093/nar/gkz1021>.
- Pomaznoy, M., Ha, B., & Peters, B. (2018). GOnet: A tool for interactive Gene Ontology analysis. *BMC Bioinformatics*, 19(1), 470. Available from <https://doi.org/10.1186/s12859-018-2533-3>.
- Raudvere, U., et al. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1), W191–W198. Available from <http://doi.org/10.1093/nar/gkz369>.
- Sealfon, R. S. G., et al. (2006). GOLEM: An interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, 7(1), 443. Available from <https://doi.org/10.1186/1471-2105-7-443>.
- Straiton, J., et al. (2019). From Sanger sequencing to genome databases and beyond. *BioTechniques*, 66(2), 60–63. Available from <https://doi.org/10.2144/btn-2019-0011>.
- Subramanian, A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. Available from <https://doi.org/10.1073/pnas.0506580102>.
- Supek, F., et al. (2011). REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS One*, 6(7), e21800. Available from <https://doi.org/10.1371/journal.pone.0021800>.
- Sureyya Rifaioglu, A., et al. (2019). DEEPred: Automated protein function prediction with multi-task feed-forward deep neural networks. *Scientific Reports*, 9(1), 1–16. Available from <https://doi.org/10.1038/s41598-019-43708-3>.
- Tian, T., et al. (2017). AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, 45(W1), W122–W129. Available from <https://doi.org/10.1093/nar/gkx382>.

- Tokar, T., et al. (2020). GSOAP: A tool for visualization of gene set over-representation analysis. *Bioinformatics*, 36(9), 2923–2925. Available from <https://doi.org/10.1093/bioinformatics/btaa001>.
- Torto-Alalibo, T., Collmer, C. W., & Gwinn-Giglio, M. (2009). The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: Community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. *BMC Microbiology*, S1. Available from <https://doi.org/10.1186/1471-2180-9-S1-S1>.
- Wei, Q., et al. (2017). NaviGO: Interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *BMC Bioinformatics*, 18(1), 177. Available from <https://doi.org/10.1186/s12859-017-1600-5>.
- Xu, Q., & Shaulsky, G. (2005). GOAT: An R tool for analysing gene ontology™ term enrichment. *Applied Bioinformatics*, 4(4), 281–283. Available from <https://doi.org/10.2165/00822942-200504040-00008>.
- Xuan, P., et al. (2019). Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Frontiers in Genetics*, 10(MAY). Available from <https://doi.org/10.3389/fgene.2019.00459>.
- Yang, J., et al. (2016). RNA-seq transcriptome analysis of a *Pseudomonas* strain with diversified catalytic properties growth under different culture medium. *MicrobiologyOpen*, 5(4), 626–636. Available from <http://doi.org/10.1002/mbo3.357>.
- Yang, Y., et al. (2018). Bioinformatic identification of key genes and pathways that may be involved in the pathogenesis of HBV-associated acute liver failure. *Genes and Diseases*, 5(4), 349–357. Available from <https://doi.org/10.1016/j.gendis.2018.02.005>.
- Ye, J., et al. (2018). WEGO 2.0: A web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research*, 46(W1), W71–W75. Available from <https://doi.org/10.1093/nar/gky400>.
- Yu, G., et al. (2015). DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4), 608–609. Available from <https://doi.org/10.1093/bioinformatics/btu684>.
- Zeng, X., Zhang, X., & Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Briefings in Bioinformatics*, 17(2), 193–203. Available from <https://doi.org/10.1093/bib/bbv033>.
- Zhang, J., et al. (2019). Integrating multiple heterogeneous networks for novel LncRNA-disease association inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(2), 396–406. Available from <https://doi.org/10.1109/TCBB.2017.2701379>.
- Zhao, Y., et al. (2020). A literature review of gene function prediction by modeling Gene Ontology. *Frontiers in Genetics*. Available from <https://doi.org/10.3389/fgene.2020.00400>.
- Zheng, Q., & Wang, X. J. (2008). GOEAST: A web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, 36(Web Server issue). Available from <https://doi.org/10.1093/nar/gkn276>.
- Zhou, X., & Su, Z. (2007). EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*, 8(1), 246. Available from <https://doi.org/10.1186/1471-2164-8-246>.