10

Gene Ontology, Enrichment Analysis, and Pathway Analysis

Tao Zhou, Jun Yao and Zhanjiang Liu

Introduction

Genomic sequencing has revealed that a large fraction of the genes specifying the core biological functions are conserved across a broad spectrum of eukaryotes. Knowledge of the biological role of the shared genes in one organism can often be transferred to other organisms. Shared vocabularies are an important step toward unifying biological databases and making biological knowledge transferable. Gene ontology (GO) provides a set of structured, controlled vocabularies for community use in gene annotation. With various sequence datasets including peptide sequences, genes, ESTs, microarray datasets, RNA-Seq datasets, or whole genome sequences, GO analysis provides defined GO terms to genes. At the highest level, GO terms cover cellular components, molecular functions, and biological processes. Lower levels of GO terms can be applied to genes when relevant.

While GO analysis is useful, GO terms are generally too general to provide specific insights into the mechanisms of expression changes under a specific condition. Enrichment analysis is designed to help the researchers to interpret the expression data such that, from a list of differentially expressed genes, one can determine which sets of genes are over- or under-represented, providing insights into the potential molecular basis of the biological process under study. Further, pathway analysis has the ability to determine what pathways the enriched genes are involved with, reducing the complexity of analysis while increasing explanatory power.

GO and the GO Project

Gene ontology is a controlled vocabulary term to describe gene characteristics in terms of their localization and function. It is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. The GO project (The Gene Ontology Consortium *et al*., 2000) (http://www.geneontology.org/) is considered to be the most successful example of a systematic description of biological attributes of genes in order to allow the integration, retrieval, and computation of data (Yon Rhee *et al*., 2008). GO was initially developed by researchers studying the genome of three

Bioinformatics in Aquaculture: Principles and Methods, First Edition. Edited by Zhanjiang (John) Liu.

model organisms: *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), and *Saccharomyces cerevisiae* (yeast) in 1998 (The Gene Ontology Consortium *et al*., 2000). Now, databases for many other model organisms have joined the GO Consortium, and made contributions to this project (The Gene Ontology Consortium, 2015). The GO Consortium was aimed at producing a dynamic, controlled vocabulary that could be applied to all eukaryotes. The GO project provides three structured ontologies that describe gene products in terms of their biological processes, cellular components, and molecular functions in a species-independent manner.

GO Terms

Each GO term within GO has a term name, which may be a word or string of words; a unique alphanumeric identifier; a definition with cited sources; or a namespace indicating the domain to which it belongs. Terms may also have synonyms, which are classified as being exactly equivalent, broader, narrower, or related to the term name. The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, and single-cell and multicellular organisms.

Ontology

Ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of terms with relations that operate between them. The domains that GO represents are biological processes, molecular functions, and cellular components.

Biological Process

Biological process refers to a biological objective to which the gene or gene product contributes. A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "GO:0044699 single organism process" or "GO:0032502 developmental process". Examples of more specific terms are "GO:0042246 tissue regeneration" or "GO:0031101 fin regeneration". The general rule to distinguish between a biological process and a molecular function is that a process must have more than one distinct step. We should also be aware that a biological process is not equivalent to a pathway.

Molecular Function

Molecular function is defined as the biochemical activity of a gene product, such as binding to specific ligands. Molecular function terms describe activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. For instance, "GO:0042813 Wnt-activated receptor activity" is a molecular function term.

Cellular Component

Cellular component refers to the place in the cell where a gene product is active. Cellular component terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g., rough endoplasmic reticulum or nucleus) or a

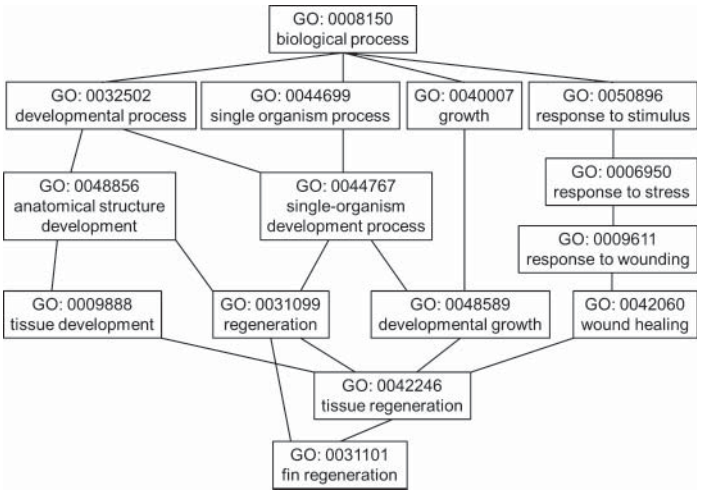

Figure 10.1 Schematic presentation of an example of the ontology structure.

gene product group (e.g., ribosome, proteasome, or a protein dimer). For instance, "GO:0009898 cytoplasmic side of plasma membrane" is a cellular component term.

Ontology Structure

The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains. For instance, ribosomal protein S2 is a component of the cellular component ribosome, it is involved in the translation process, and it binds to RNA. Each GO term is a node, and the relationships between the terms are edges between the nodes. Ontology structure is loosely hierarchical, with "child" terms being more specialized than their "parent" term. A GO term may have more than one parent term. Figure 10.1 shows a set of terms from the ontology. GO terms do not occupy strict fixed levels in the hierarchy. GO is structured as a graph, and terms would appear at different "levels" if different paths were followed through the graph.

GO Slim

GO slims are subsets of terms in the ontology. GO slims give a broad overview of the ontology content without the details of the specific fine-grained terms. GO slims are very useful when providing a summary of the results of the GO annotation of a genome, microarray, or cDNA collection where broad classification of gene product functions is required. GO slims are created according to the user's needs, and may be specific to species or to particular areas of the ontologies. GO also provides a generic GO slim that, as with GO itself, is not species-specific, and should be suitable for most purposes. The GO Consortium website provides GO slims related to different organisms or usages in Open Biomedical Ontologies (OBO) format for download (http://geneontology.org/page/go-slim-and-subset-guide).

Annotation

GO annotation is the process of assigning GO terms to gene products. Genome annotation is the practice of capturing data about a gene product. Although genome annotation provides annotation of genes, it is usually less detailed than GO annotations. The detailed information of a gene product can be found by GO annotations. Several computational methods can be used for mapping GO terms to gene products. The GO Consortium recommends three annotation processes: electronic annotation, literature annotation, and sequence-based annotation.

Electronic Annotation

The vast majority of GO annotations have been made using electronic annotation methods, without curators' oversight. Electronic annotation is very quick and produces large amounts of less detailed annotation. Electronic annotation is likely to tell you which of your genes are transcription factors, but unlikely to tell you in great detail what process the gene controls. Electronic annotation is especially useful for the annotation of new genomes or microarrays with thousands of sequences. However, it has been reported that electronic annotation is more reliable than generally believed, and that it can even compete with annotation inferred by curators when they use evidence other than experiments from primary literature (Škunca, Altenhoff & Dessimoz, 2012).

The primary method of generating electronic annotation is to manually map GO terms to corresponding concepts in the controlled vocabularies. Electronic annotations can be inferred from Enzyme Commission, InterPro, UniProt, Ensembl, HAMAP, BLAST, etc. All of these methods involve the use of mapping files, which can be downloaded from the GO Consortium website (http://geneontology.org/page/download-mappings). These mapping files are updated regularly, and can be downloaded in the ".txt" file format or viewed in a web browser. The line syntax for the mapping file is: "external database:term identifier (id/name) > GO:GO term name ; GO:id". Each line represents one mapping item. One external database term may direct to many GO terms.

InterPro provides the functional analysis of proteins by classifying them into families and predicting functional domains and important sites (Mitchell *et al*., 2015). InterPro provides mapping of InterPro entries to GO terms, and can be downloaded from the InterPro website (www.ebi.ac.uk/interpro/download.html). For example: "InterPro:IPR000003 Retinoid X receptor/HNF4 > GO:DNA binding ; GO:0003677".

The UniProt Knowledgebase (UniProtKB) is a collection of functional information on proteins, with accurate, consistent, and rich annotations. UniProtkb captures the core data and as much annotation information as possible for each UniProtkb entry. UniProt keyword mapping, generated by the UniProtKB and UniProtKB-GOA teams, is aimed at assigning GO terms to UniprotKB keywords (Barrell *et al*., 2009). For example: "UniProtKB-KW:KW-0067 ATP-binding > GO:ATP binding ; GO:0005524".

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource (www.genome.jp/kegg/) for understanding high-level functions and utilities of the biological system, such as cells, organisms, and ecosystems, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. The detailed information for each KEGG entry, such as definition, enzyme, pathway, and ortholog, can be retrieved from the KEGG website. KEGG mapping is useful for subsequent pathway analysis. KEGG

pathways and reactions mapping can map GO terms to KEGG entries. For example: "KEGG:R00004 > GO:inorganic diphosphatase activity ; GO:0004427".

Electronic annotations can also be made by aligning your own sequences to manually annotated sequences, then transferring the GO annotations across to your sequences. This is quite useful for huge amounts of sequences. The threshold of similarity or E-value in this process can be controlled. Many mapping files for different databases can be found on the GO Consortium website. All of the mapping files have the same format and usage. You may download the required mapping file and import it into Microsoft Access. By querying your data with the mapping file in Microsoft Access, it is easy to annotate large amounts of data.

Literature Annotation

Literature annotation in GO means capturing published information about the exact function of gene products as GO annotations. Literature annotation usually starts with a list of genes and is conducted through the following processes.

1) Read the publication about the gene and collect useful information such as function, biological process, and cellular location.
2) Search the NCBI taxonomy database (www.ncbi.nlm.nih.gov/taxonomy) and find the taxon id of the organism that the gene product is derived from (Format-Taxonomy ID: 7955).
3) Find the accession number of the gene product, for example, from UniProt, DDBJ, or GeneBank.
4) Find the GO terms that describe the function, biological process, or cellular location from the GO database (http://amigo.geneontology.org/amigo).
5) Choose the evidence code—for example, Inferred from Experiment (EXP)—that describes the experiment. The guide to GO evidence codes can be found in the Go Consortium website (http://geneontology.org/page/guide-go-evidence-codes).
6) The information can be submitted to the GO Consortium in the Gene Association File (GAF) format or the Gene Product Association Data (GPAD) format. GO annotation file formats can be found from the GO Consortium website (http://geneontology.org/page/go-annotation-file-formats).

Literature annotation is time-consuming, but produces high-quality annotation, and is worthwhile in the long term.

Sequence-Based Annotation

Researchers may start with some sequences, and want to assign GO terms to them. In this scenario, sequence-based annotation can be performed. This process consists of the following steps:

1) *Blast search*. Once a sequence has been selected for annotation, Blast searches can be conducted against UniProtKB to capture genes related to it and to identify homologs. The sequence and homologs can be merged into a single entry. The difference about the homologs, such as alternative splicing, natural variations, and frameshifts, can be documented for further comparison.
2) *Sequence analysis*. Sequence annotation prediction processes can be found in Chapter 4. The relevant gene, domain, protein topology, and protein family classification can be predicted after sequence analysis.

3) *Literature review*. The publications related to the gene can be identified by searching literature databases. Read the publications and collect information about the gene, such as function, biological process, or cellular location. The information collected through literature review and sequence analysis tools can be compared and verified. Annotation captured from literature review includes protein and gene name, function, catalytic activity, subcellular location, protein–protein interactions, patterns of expression, disease associated with deficiencies in a protein, RNA editing, etc. Relevant GO terms are assigned based on the experimental data documented in the literature.
4) *Family-based analysis*. Reciprocal Blast searches and phylogenetic resources (discussed in Chapter 2) can be used to identify putative homologs. Annotation is standardized and propagated across homologous proteins to ensure data consistency.
5) *Evidence attribution*. All information added to an entry during the manual annotation process should be linked to the original source, so that researchers can trace back to the origin of each piece of information and evaluate it. Evidence code can be assigned following the GO Consortium guidelines (http://geneontology.org/page/guide-go-evidence-codes).
6) *Quality assurance, integration, and update*. Each completed entry can be submitted to the GO Consortium in the GAF or GPAD format. After quality assurance, the complete entry will be updated and becomes available.

GO Tools

The GO Consortium develops and supports AmiGO 2, OBO-Edit, and GOOSE. There are a large number of third-party tools available that use the data provided by the GO project.

AmiGO 2

AmiGO 2 (http://amigo.geneontology.org/amigo) is a web-based application that allows users to query, browse, and view ontologies and gene product annotation data. AmiGO 2 can be used online at the GO website to access the data provided by the GO Consortium, or can be downloaded and installed for local use (only supports the Ubuntu platform). The AmiGO 2 manual can be found on the GO Consortium website (http://wiki.geneontology.org/index.php/Category:AmiGO_2_Manual). In the following text, we will introduce the usage of AmiGO 2.

Quick Search AmiGO 2 quick search is a powerful method that rapidly searches across all of the pre-computed indexes. Quick search supports Boolean operator, wildcard, and fuzzy searches. As you type in the search box, the autocomplete search will find matching terms and gene products. Once a term or gene product has been selected from the dropdown list, it will jump directly to that selection's detail page. Otherwise, by pressing return or clicking the search button, the user can jump to an annotation search for his or her text. Boolean logic and nesting may be used in the search. The following are some examples for quick search.

1) If you want to search the records that contain both angiogenesis and neurogenesis, you would enter: "angiogenesis and neurogenesis".

2) If you want to exclude neurogenesis from angiogenesis results, you would enter: "angiogenesis and – neurogenesis".
3) If you want to get all the results for angiogenesis, neurogenesis, or both, you would enter: "angiogenesis or neurogenesis". In this case, "angiogenesis or neurogenesis" and "angiogenesis neurogenesis" are functionally equivalent. Space between words are considered to be an implicit "or".
4) You need to use parentheses to club together words that you want to appear together. For example, if you want to search for records containing both "angiogenesis" and "wound healing", you would enter: "angiogenesis and (wound healing)".
5) If you want to search for "angiogenesis" in conjunction with either "neurogenesis" or "organogenesis", you would enter: "angiogenesis and (neurogenesis or organogenesis).

Advanced Search Advanced search can be used to search the GO database interactively. To start, click the search button, and select "annotations", "ontology", or "gene and gene products" from the drop-down list.

The following text lists the steps involved in searching for annotations. In this example, we will try to find the tgf-beta family proteins involved in angiogenesis:

1) Click "annotations" from the search button drop-down list.
2) Type "angiogenesis" into the free-text filtering box.
3) Click "PANTHER family" from the filters list.
4) Select "tgf-beta family".
5) From the found entities, you may browse the results.
6) More filters such as "source", "assigned by", "ontology", "evidence type", etc., can be used to narrow the results.

The following text lists the steps involved in searching for ontology. In this example, we will try to explore blood vessel development:

1) Click "ontology" from the search button drop-down list.
2) Type "blood vessel development" into the free-text filtering box.
3) Click "ontology source" from the filters list, and select "biological process".
4) Click "ancestor" from the filters list, and select "tissue development".
5) From the found entities, you may browse the results.

The following text lists the steps involved in searching for genes and gene products. In this example, we will try to find the genes and gene products in wnt family involved in fin regeneration:

1) Click "genes and gene products" from the search button drop-down list.
2) Type "fin regeneration" into the free-text filtering box.
3) Click "PANTHER family" from the filter list, and select "wnt related".
4) From the found entities, you may browse the results.

Grebe Grebe is designed for users who are unfamiliar with "quick search" and "advanced search", described in the preceding text. The Grebe search wizard can be used to quickly answer common questions using a fill-in-the-blanks approach. The use of the Grebe search wizard is straightforward: fill in the necessary blank columns with IDs, terms, or

gene products; select from the autocomplete results; and clink the "GO" button at the end of the question.

GOOSE

GO Online SQL Environment (GOOSE) is a web environment that allows users to freely enter SQL queries to the GO database. Users familiar with SQL can enter queries into the box. After selecting one of the available mirrors, click the "Query" button. For users who are new to SQL, the GO Consortium provides many sample queries. Users can select the sample query that meets their needs, and substitute the keywords.

In this example, we will try to find all genes directly annotated to "angiogenesis" (excluding child terms):

1) Click the button "Go" under GOOSE in the main page of AmiGO 2, or visit the website (http://amigo.geneontology.org/goose) to enter the GO online SQL environment.
2) From the drop-down list of "use an example query", select "All genes directly annotated to 'nucleus' (excluding child terms)". You may notice that queries are automatically filled in the form field under "Directly query GO data using SQL".
3) Substitute the value of "term.name" from "nucleus" to "angiogenesis" at the end of the query.
4) Select one mirror that you believe will perform well for your needs from the available mirrors. Mirrors may have different load settings and frequencies.
5) Select the number of results you want to be returned.
6) Check the box "Download results directly in a text format" if you want to download the results.
7) Click the query button, and you will find the results.

Although many sample queries have been listed in GOOSE, users may have special or complex queries. The GO Consortium has listed various kinds of possible queries on the GO LEAD database in the GO wiki main page (http://wiki.geneontology.org/index.php/Example_LEAD_Queries). Users may develop their own queries by modifying these sample queries.

Blast2GO

Blast2GO (Conesa *et al*., 2005) is a bioinformatics platform for the functional annotation and analysis of genomic datasets. Blast2GO provides a user-friendly interface for GO annotation. Users may design custom annotation styles through the many configurable parameters. Blast2GO does not only generate functional annotations, it also supports InterPro domains, RFAM IDs, enzyme codes (ECs), and KEGG maps. Additionally, Blast2GO provides a wide array of graphical and analytical tools for annotation manipulation and data mining. There are two editions of Blast2GO: basic and professional. Blast2GO basic is a free and simplified edition, with a limited number of databases and basic functions. Blast2GO PRO is the commercial edition, which provides more databases, parameters, and functions.

A typical analysis process of Blast2GO consists of five steps: BLAST, mapping, annotation, statistical analysis, and visualization. Basically, Blast2GO uses BLAST searches to find sequences similar to the input sequences. Mapping is the process of extracting the GO terms and ECs associated with each of the obtained hits. Annotation is an

evaluation of the extracted GO terms. GO annotation can be viewed by reconstructing the structure of the GO relationships and ECs highlighted on the KEGG maps. Here, we will provide a general description for the usage of Blast2GO:

Blast2GO installation: Blast2GO can be downloaded from its website, (https://www.blast2go.com/blast2go-pro/download-b2g) and installed on Microsoft Windows, Mac, and Linux platforms. At least 1 GB of RAM and a working Internet connection are required for the application. Users may register for a free Blast2GO basic account or subscribe to Blast2GO PRO to activate the software.

Blast2GO user interface: There are four basic sections in the Blast2GO main user interface—the menu bar, the main analysis icons, the main sequence table, and the application tab.

General usage: Here, we provide an introduction to the general usage of Blast2GO. Users may find more detailed information from the Blast2GO manual (https://www.blast2go.com/images/b2g_pdfs/b2g_user_manual.pdf).

1) *Load sequence*. Click "File" - > "Load" - > "Load Sequence" and select the ".fasta" file containing the set of sequences. Blast results in xml format can also be loaded.
2) *Blast*. Clicking on the "Blast" icon will initiate the processes. At the Blast configuration dialog, select the Blast mode that is appropriate for your sequence type. You may select "blastx" for nucleotide and "blastp" for protein data. Click "Next" for advanced settings and choose the location to save the Blast results. Click "Run" to start the Blast search against an NCBI non-redundant (nr) database. Once the Blast analysis is completed, you may view the results by clicking the "Chart" icon -> "Blast statistics". On the main sequence table, right-click on a sequence and select "Show sequence result" to open the Blast results for the single sequence. The sequences with Blast hits will turn into orange, and those without Blast hits will turn into red.
3) *InterProScan*. Click on the "Interpro" icon, and the corresponding wizard will be shown. You may need to fill in your email address and choose applications from the wizard's list. Click "Run" after you have selected where to save the results. InterProScan can be run in parallel with Blast.
4) *Mapping*. Click on the "Mapping" icon and click "Run" to start mapping process. The mapping process will associate the Blast hits of each sequence with GO terms. Successfully mapped sequences will turn green. Mapping results can be viewed by clicking the "Charts" icon -> "Mapping Statistics".
5) *Annotation*. Clicking on the "Annot" icon will start the annotation configuration wizard. By clicking "Next", you will start the evidence code weights wizard. Change the parameters if needed and click "Run" to start the annotation. Successfully annotated sequences will turn blue. Mapping results can be viewed by clicking the "Charts" icon -> "Annotation Statistics".
6) *Enrichment analysis*. Blast2GO provides enrichment analysis for the statistical analysis of GO term frequency differences between two sets of sequences. Clicking "Analysis" - > "Enrichment Analysis (Fisher's exact test)" will open Fisher's exact test configuration page. Select a ".txt" file containing a sequence ID list for a subset of sequences. You may set all the loaded sequences as references, or upload a second set of files containing the reference sequences. Click on "Run" button to start the analysis. A table containing the results of the enrichment analysis will be displayed. Click on the "Make Enriched Graph" icon to view the results of the Fisher's exact

test on the GO DAG. Click on "Show Bar Chart" to obtain a bar chart representation of the GO term frequencies.

7) *Combined graph*. Blast2GO provides visualization of the combined annotation for a group of sequences. Select a group of sequences to generate their combined graph. The "Select by color" option under the "Select" icon may help in selecting the featured sequences quickly. Users can also select the sequences manually by clicking the checkbox near each sequence. After the sequences are selected, click "Graphs" -> "Make combined Graph", select the functional annotation to be viewed, and click "Run" to generate the graph.
8) *Save results*.
   Clicking "File" -> "Save" saves the current Blast2GO project as a ".b2g" file.
   Clicking "File" -> "Export" allows users to export various results in different formats.
   The Fisher's exact test results can be saved by clicking "Save as text".
   The graphs can be saved by clicking "Save as" in the toolbar located on the right side of the graphs.

Enrichment Analysis

Researchers usually get sets of differentially expressed genes after performing high-throughput experiments such as RNA-Seq and microarrays. Although such information is useful, just generating a list of genes ("listomics") is not insightful for biology. One may wish to retrieve a functional profile of the gene sets for a better understanding of the underlying biological processes. The comparison of differentially expressed gene sets with the gene sets involved in most biological processes may provide at least some insights into the involved biological mechanisms. This can be achieved by enrichment analysis, which compares the input gene set with the reference to determine if it is enriched. Gene set enrichment or functional enrichment analysis is a method to identify classes of genes or proteins that are over-represented in a set of genes or proteins (Subramanian *et al*., 2005). The method uses statistical approaches to identify significantly enriched or depleted groups of genes. The principal foundation of enrichment analysis is that a gene set should have a higher chance to be selected if its underlying biological process is abnormal in a given study (Huang, Sherman & Lempicki, 2009).

Main Types of Enrichment Tools

There has been an explosion in the number of software packages available for annotation enrichment analysis. About 68 tools were uniquely categorized into three major classes according to their underlying enrichment algorithms: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA) (Huang *et al*., 2009). Some tools with different capabilities belong to more than one class.

SEA is the most traditional enrichment approach that iteratively tests annotation terms one at a time against a list of users' pre-selected genes for enrichment. Enriched annotation terms that pass the enrichment p-value threshold are reported. SEA uses a simple strategy, which is very efficient in extracting the major biological meaning

behind large gene lists. Most of the earlier tools used this strategy, and proved to be significantly successful in many genomic studies. However, a common weakness for this class of tools is that the linear output of terms can be very large and overwhelming.

GSEA adopts a "no-cutoff" strategy that takes all genes from an experiment without selecting significant genes. GSEA reduces the arbitrary factors in the gene selection step and uses all information obtained from the experiments. GSEA methods need biological values such as fold changes for each of the genes as input. It is sometimes difficult to summarize many biological aspects of a gene into one value when the experiment design or genomic platform is complex.

MEA considers the relationships existing between different annotation terms during enrichment. The GO structure is loosely hierarchical, in which some joint terms may contain unique biological meanings. Researchers can take advantage of term-to-term relationships by using the MEA method, which can also reduce redundancy and prevent the dilution of potentially important biological concepts. However, the disadvantage of MEA is that some terms or genes with weak relationships to neighboring terms or genes could be left out from the analysis.

Gene Set and Background

The input files for the three types of enrichment tools are different. SEA is a list-based method that needs a subset of all genes chosen by some relevant method and a list of annotations linked to genes. GSEA is a rank-based method. The inputs for GSEA are a set of all genes ranked by some metrics such as fold change, and a list of annotations linked to genes. The MEA method is also list-based, but needs term-to-term relationships. The MEA method requires a subset of all genes, and a list of annotations linked to genes that are organized in some relationship.

Defining the background is very important in enrichment methods. To get the real meaning of enrichment with respect to the experiment, researchers may need to upload the reference. For example, if you are looking at a set of genes from a particular tissue, a reference for that tissue would allow generation of more meaningful results than a reference for the whole genome.

Annotation Sources

Many different annotation categories can be used by enrichment analysis, including biological function (e.g., GO terms), physical position (e.g., chromosomal location), regulation (e.g., co-expression), protein domains, pathways, or other attributes for which prior knowledge is available. GO database is very suitable for enrichment analysis on gene sets. GO has become one of the most popular annotation sources.

Statistical Methods

The most popular statistical methods used in enrichment calculation are Fisher's exact test, chi-square test, hypergeometric distribution, and binomial distribution (Huang *et al*., 2009). Binomial probability is believed to be suitable for analysis with a large population background on a principal level. Fisher's exact test, chi-square test, and hypergeometric distribution are better for analysis with a smaller population background (Khatri & Drăghici, 2005). Each statistical method has its own weakness and limitations. It is not

realistic to choose enrichment analysis tools simply according to statistical methods. Different tests may produce very different ranges of p-values. Users may try different statistical methods on the same dataset and compare the results whenever possible.

Recommended Tools

It may be difficult for users to select suitable tools from the long list of available tools. It is better to choose tools that include the species that you are researching. Make sure the tools accept your input identifiers and have the most updated annotations. Try several tools to determine the one that best fits your study purpose. Database for annotation, visualization, and integrated discovery (DAVID; Huang, Sherman & Lempicki, 2008), protein analysis through evolutionary relationships (PANTHER; Mi, Muruganujan & Thomas, 2013), and GSEA (Subramanian *et al*., 2005) are all very good tools to start with. In the following text, we will introduce DAVID.

Enrichment Analysis by Using DAVID

DAVID (Huang *et al*., 2008) is a web-accessible program (https://david.ncifcrf.gov/) that provides a comprehensive set of functional annotation tools. It highlights the most relevant GO terms associated with a given gene list, and belongs to the MEA method. DAVID has over 40 annotation categories, including GO terms, protein–protein interactions, protein functional domains, disease associations, bio-pathways, sequence general features, homologies, gene functional summaries, gene tissue expressions, and literature. The extended annotation coverage provides researchers with much power to analyze their genes for many different biological aspects. DAVID accepts customized gene background, which more specifically meets user requirements for the best analytical results.

Universal Gene List Manager Panel The gene list manager panel is centralized and universal for all DAVID tools. Users can upload gene lists or backgrounds through this panel, so that different DAVID tools can access them. Users do not need to re-submit the same gene list for different DAVID tools.

The following are the specifics of the gene list manager panel's "Upload" tab:

1) *Enter gene list*. Users may paste the gene list into the box or upload it from a file. The gene list should be listed in the format of one gene per row, without a header row. DAVID is case-insensitive for all the accessions or IDs.
2) *Select identifier*. From the drop-down box, select the corresponding gene identifier type for the genes in the pasted/uploaded gene list. For users who are not sure about the identifier type, the "not sure" option can be selected. After submitting, users will be led to the gene ID conversion tool, where the possible source of gene IDs can be analyzed. Users may go back to select the identifier, or convert the gene ID, and then submit to DAVID as a gene list or background.
3) *Select list type*. If users are uploading gene lists for annotation analysis—for example, selected differentially expressed genes from the RNA-Seq experiment—the "Gene list" option can be selected. "Background" means the submitted genes as a customized gene background are for enrichment background calculation purpose—for example, the entire genes in an array. The customized background is useful when all the pre-built backgrounds in DAVID do not satisfy the user's particular purpose. The gene list submitted as a background will show up in the "Background" tab.

4) *Submit*. Click "Submit list" to submit the list to DAVID. You should see the corresponding gene lists listed in the "List" tab or "Background" tab if the submission is successful. Moreover, an expected gene number should also be associated with the gene lists.

The following are the specifics of the gene list manager panel's "List" tab:

1) The species information is listed in the "List" tab. Users may choose to limit annotations by one or more species to analyze together or separately from the top box if the gene list contains multiple species.
2) Click the "Select" button to switch species.
3) From the bottom box (list manager), users may highlight the gene list to be analyzed.
4) Click the "Use" button to switch gene list. Users may also rename, remove, combine, or show gene list by clicking each of the corresponding buttons.

The following are the specifics of the gene list manager panel's "Background" tab:

The DAVID default backgrounds and user-uploaded backgrounds are listed in the top box. Users may highlight a background, and click "Use" to switch backgrounds. There are many pre-built array backgrounds listed at the bottom part of the population manager. Users can click the checkbox near an array background to switch. The successfully switched background will be shown in "Current background" under "Annotation summary results".

The selection of a population background will affect results significantly. However, there is no background to fit all the situations of various studies. The DAVID default population background in enrichment calculation is the corresponding genome-wide genes with at least one annotation in the analyzing categories. For studies with genome-wide scope or close to genome-wide scope, the default background is a good choice. For a gene list derived from Affymetrix microarray or Illumina studies, the Affymetrix chip and Illumina chip backgrounds will be better choices, respectively. For studies far below genome-wide scope, a customized background may be a better choice.

Functional Annotation Tool The functional annotation tool mainly provides batch annotation and enrichment analysis. Functional annotation clustering uses a novel algorithm to measure relationships among annotation terms that reduce the burden of associating similar redundant terms and make the biological interpretation more focused at a group level. The functional annotation tool can display genes from a user's list on pathway maps to facilitate biological interpretation in a network context.

1) *Load gene list*. Users can load gene lists and references using to the universal gene list manager.
2) *View annotation summary results*. The selected gene list and background will be shown at the top of the annotation summary results. Users may view and select annotation categories from the middle part of the summary. The combined view for selected annotations will be presented by clicking the corresponding toolbar.
3) *Explore details using the functional annotation clustering report, chart report, and table report*. The functional annotation clustering report groups similar annotations together. The grouping algorithm is based on the hypothesis that similar annotations would have similar gene members. The functional annotation chart report lists annotation terms and their associated genes. Fisher's exact test statistics is calculated

based on the corresponding DAVID gene IDs. Options can be set to ensure the display of only statistically significant genes. In comparison, the functional annotation table report lists the genes and their associated annotation terms. There is no statistics applied in the annotation table report.

4) *Export and save results*. The functional annotation clustering report, chart report, and table report can be downloaded in ".txt" file format by clicking the "Download file" option in each of the reports.

Gene Functional Classification The gene functional classification tool generates a gene-to-gene similarity matrix based on shared functional annotations. The gene similarity matrix can systematically enhance biological interpretations of large lists of genes derived from high-throughput experiments. The functional classification tool provides a rapid means to organize large lists of genes into functionally related groups to help unravel the biological content captured by high-throughput technologies.

Users can click "Gene functional classification" in "Shortcut to DAVID tools menu" to initiate the tool. Then, the gene list needs to be submitted to the DAVID system. Users can then view the results in text mode or in heat map view.

Gene Accession Conversion Tool The gene accession conversion tool can be used to convert gene lists to identifiers of the most popular resources. The given gene accession can be quickly mapped to another, based on the user's choice. Suggestions of possible choices for ambiguous gene accessions in gene lists can also be automatically provided by this tool.

1) Select genes in the gene list and the corresponding species to be converted.
2) Select the final gene identifier type to be converted to.
3) Click submit.

After submission, the result page will be shown. The genes that have been successfully converted to the desired identifiers will be displayed in the right panel. The statistics summary of gene accession conversion will be displayed in the left panel. Users may download the list or submit the converted list to DAVID directly.

Gene Pathway Analysis

Genes do not function individually independently. A gene may be one of the many genes involved in a specific gene pathway. Given a list of genes involved in biological processes, researchers may wish to map them to known pathways and determine which pathways are over-represented in a given set of genes. Pathway analysis allows reduction of complexities and increases in explanatory power. It has become a good choice for gaining insight into the underlying biology for a given gene or protein list. In the following text, we discuss the different generations of pathway analysis tools and list some pathway analysis databases.

Definition of Pathway

Different biological definitions of *pathways* are used in different pathway databases. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway combines multiple biological processes from different organisms to produce a substrate-centered reaction

mosaic. The BioCyc ontology defines a pathway as a conserved, atomic module of the metabolic network of a single organism (Green & Karp, 2006). Different pathway concepts can lead to different outcomes from a computational study that relies on pathway databases.

The knowledge-driven pathway analysis refers to the methods that exploit pathway knowledge in public repositories such as GO or KEGG, rather than to methods that infer pathways from molecular measurements (Khatri, Sirota & Butte, 2012).

Pathway Analysis Approaches

The evolution of knowledge-driven pathway analysis has been distinctly divided into three generations, each with its advantages and disadvantages. For a comparison of pathway analysis tools for each generation, readers are referred to Khatri *et al*. (2012).

Over-representation Analysis (ORA) Approaches

ORA statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression. ORA uses the following strategy (Khatri *et al*., 2012): First, create an input gene list using a certain threshold, such as differentially expressed genes. Then, count the genes involved in each pathway. Next, every pathway is tested based on the hypergeometric, chi-square, or binomial distribution for over- or under-representation in the list of input genes.

ORA has some limitations, despite that it is a widely used tool. ORA treats each gene equally, ignoring the correlation structure between genes and useful information about the extent of regulation, such as fold change. ORA may result in information loss by using only the most significant genes. Also, ORA assumes that each pathway is independent from others, which is erroneous.

Functional Class Scoring (FCS) Approaches

FCS hypothesizes that weaker but coordinated changes in sets of functionally related genes can have significant effects on pathways. FCS consists the following steps (Khatri *et al*., 2012): First, gene-level statistics for the differential expression of individual genes is generated using the molecular measurements from an experiment. Second, the gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic. Finally, the statistical significance of the pathway-level statistic is assessed. FCS overcomes some limitations of ORA, but it also has some limitations. Similar to ORA, FCS analyzes each pathway independently, which ignores the cross and overlap among pathways. FCS methods use changes in gene expression to rank genes in a given pathway, and discard the changes from further analysis.

Pathway Topology (PT)–based Approaches

PT-based methods have been developed to utilize additional information about gene products, including interaction with each other in a given pathway. PT-based methods are essentially the same as FCS methods, except for the use of pathway topology to compute gene-level statistics. They are improved from FCS methods, but also have several common limitations. True pathway topology is rarely available and is fragmented in the knowledge base. PT-based methods are unable to model the dynamic states of a system, and cannot consider interactions between pathways (Khatri *et al*., 2012).

Pathway Databases

Pathway information is available through a large number of databases. These databases can display pathway diagrams, which combine metabolic, genetic, and signal networks. Most databases are created by extracting pathway information from journal articles and then organizing the information with pathway diagrams. It is necessary for users to select proper pathway databases for their research. In the following text, we introduce some of the major pathway databases.

KEGG

KEGG (Kanehisa & Goto, 2000) (http://www.genome.jp/kegg/) is an encyclopedia that collects all knowledge relevant to biological systems. KEGG pathway is a collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human disease, and drug development. Users can map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretation of higher-level systemic functions.

Reactome

Reactome (Joshi-Tope *et al*., 2005) (http://www.reactome.org/) is a curated database of pathways and reactions (pathway steps). The goal of Reactome is to provide intuitive bioinformatics tools for the visualization, interpretation, and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology, and education. Information in the database is authored by expert biologist researchers, maintained by Reactome editorial staff, and extensively cross-referenced to other resources—for example, NCBI, Ensembl, UniProt, UCSC Genome Browser, HapMap, KEGG, ChEBI, PubMed, and GO.

PANTHER

PANTHER pathway (http://www.pantherdb.org/pathway/) is a part of the PANTHER protein classification system (Mi *et al*., 2005). The PANTHER pathway database consists of over 177 signaling and pathways. These primary signaling and pathways have subfamilies and protein sequences mapped to individual pathway components. A pathway component is usually a single protein in a given organism, but multiple proteins can sometimes play the same role. PANTHER pathways capture molecular-level events in both signaling and metabolic pathways. Pathway diagrams are interactive and include tools for viewing gene expression data in the context of the diagrams. The PANTHER pathway database aims to comprehensively represent biological knowledge concerning both metabolic and signaling pathways, and to provide structured visualization of pathways for biological and informatic analysis.

Pathway Commons

Pathway Commons (Cerami *et al*., 2011) (http://www.pathwaycommons.org) is a collection of publicly available pathway information from multiple organisms. Pathway Commons provides a comprehensive collection of biological pathways from multiple sources represented in a common language for gene and metabolic pathway analysis.

Pathways in Pathway Commons include biochemical reactions, complex assembly, transport and catalysis events, physical interactions involving proteins, DNA, RNA, small molecules and complexes, gene regulation events, and genetic interactions involving genes. Researchers can search, view, and download information from Pathway Commons.

BioCyc

BioCyc (Caspi *et al*., 2014) (http://biocyc.org/) databases provide reference on the genomes and metabolic pathways of sequenced organisms. BioCyc databases are generated by predicting the metabolic pathways of completely sequenced organisms, predicting the genes codes for missing enzymes in metabolic pathways, and predicting operons. BioCyc also integrates information from other bioinformatics databases, such as protein feature and GO information from UniProt. The BioCyc website provides a suite of software tools for database searching and visualization, for omics data analysis, and for comparative genomics and comparative pathway analysis.

Pathway Analysis Tools

Many pathway analysis tools are available. We have discussed the three generations of knowledge-driven pathway analysis that may be useful for users to select proper pathway analysis tools. In the following text, we introduce several popular pathway analysis tools.

Ingenuity Pathway Analysis (IPA)

IPA (http://www.ingenuity.com/products/ipa) is the software used to display pathway data from Ingenuity Knowledge Base by QIAGEN. IPA has been broadly adopted by the life science research community, and IPA licenses require a fee. For a given gene set, IPA automatically generates the pathways that are related to those genes.

KEGG Pathway Mapping

KEGG pathway mapping (http://www.genome.jp/kegg/pathway.html) is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for the biological interpretation of high-level systemic functions. Pathways can be searched from the KEGG pathway main page by using the KEGG pathway ID or pathway name. The pathway ID may be either a reference pathway ID (beginning with "map") or a species-specific pathway ID (beginning with the three-letter organism code). For example, the MAPK signaling pathway ID is "map04010", whereas the human MAPK signaling pathway ID is "hsa04010".

To search for a specific gene, the gene ID prefixed by the species code can be searched through the KEGG table of contents. Users may also search for a specific gene in the KEGG genes database (http://www.genome.jp/kegg/genes.html). For example, typing "fgf11" in the "Search genes" textbox and clicking "Go" will search for "fgf11" in the KEGG genes database. Clicking the result that fits your requirement will display the specific gene item. For example, if we click on "has:2256", information related to the human fgf11 gene, such as definition, pathway, and sequence, etc., will be presented. There are seven pathways that fgf11 is involved in. The pathway map can be viewed by clicking the pathway ID.

PANTHER

The PANTHER gene list analysis tool (http://www.pantherdb.org/) allows users to input a list of genes (and, optionally, quantitative data) for analysis. The PANTHER batch ID searches supported IDs, including Ensembl gene identifier, Ensembl protein identifier, Ensembl transcript identifier, EntrezGene ID, Gene symbol, NCBI GI number, HGNC, IPI, UniGene, and UniProt ID. The identifiers in the user's list are automatically mapped to the primary IDs in the PANTHER database. After selecting an organism, the gene list can be analyzed in three different ways. The functional classification tool provides the classification results of the uploaded list and displays them in a gene list page or pie chart. The statistical over-representation test compares a test gene list uploaded by the user to a reference gene list and determines whether a particular class (e.g., a GO biological process or the PANTHER pathway) of genes is over-represented or under-represented. The statistical enrichment test determines whether any ontology class or pathway has numeric values that are non-randomly distributed with respect to the entire list of values (Mi *et al*., 2013). PANTHER is currently part of GO, and it integrates more updated GO curation data with the tools.

Reactome Pathway Analysis

The Reactome pathways can be viewed by the Reactome pathway browser (http://www.reactome.org/PathwayBrowser/). The "Pathways Overview" screen provides an intuitive visual overview of the Reactome hierarchical pathway structure. Most Reactome pathway topics are divided into smaller sub-pathways. Pathway diagrams show the steps of a pathway as a series of interconnected molecular events, known in Reactome as "reactions".

The Reactome pathway analysis tool (http://www.reactome.org/PathwayBrowser/#TOOL=AT) combines a number of analysis and visualization tools to permit the interpretation of user-supplied experimental datasets. Users can select the type of analysis to be performed and paste in or browse to a file containing their data. The ideal identifiers to use are UniProt IDs for proteins, ChEBI IDs for small molecules, and either HGNC gene symbols or Ensembl IDs for DNA/RNA molecules. Many other identifiers are also recognized and mapped to the appropriate Reactome molecules. After inputting the data, click "GO" to start the analysis. Analysis results are shown in the "Analysis" tab, within the "Details Panel". When you click on the name of a pathway in the "Analysis" tab, the "Pathway Hierarchy" will expand to show it, and its name will be highlighted in blue.

References

Barrell, D., Dimmer, E., Huntley, R.P. *et al*. (2009) The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, **37**, D396–D403.

Caspi, R., Altman, T., Billington, R. *et al*. (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, **42**, D459–D471.

Cerami, E.G., Gross, B.E., Demir, E. *et al*. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**, D685–D690.

Conesa, A., Götz, S., García-Gómez, J.M. *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Green, M.L. and Karp, P.D. (2006) The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, **34**, 3687–3697.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44–57.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**, 1–13.

Joshi-Tope, G., Gillespie, M., Vastrik, I. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, **33**, D428–D432.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27–30.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Khatri, P., Sirota, M. and Butte, A.J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, **8**, e1002375. doi: 10.1371/journal.pcbi.1002375

Mi, H., Lazareva-Ulitsky, B., Loo, R. *et al*. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, **33**, D284–D288.

Mi, H., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, **8**, 1551–1566.

Mi, H., Muruganujan, A. and Thomas, P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, **41**, D377–D386.

Mitchell, A., Chang, H.-Y., Daugherty, L. *et al*. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, **43**, D213–D221.

Škunca, N., Altenhoff, A. and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, **8**, e1002533. doi: 10.1371/journal.pcbi.1002533

Subramanian, A., Tamayo, P., Mootha, V.K. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.

The Gene Ontology Consortium (2015) Gene ontology consortium: going forward. *Nucleic Acids Research*, **43**, D1049–D1056.

The Gene Ontology Consortium, Ashburner, M., Ball, C.A. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Yon Rhee, S., Wood, V., Dolinski, K. and Draghici, S. (2008) Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, **9**, 509–515.