Pathway analysis tools: a comparative study over the generations

Dhusia Kalyani[1*], Rizvi Ahsan Z.[2] and Ramteke Pramod W.[3]

[1]Deptartment of Computational Biology & Bioinformatics, JSBB, SHIATS University, Allahabad-211007 (U.P.), India
[2]Indian Institute of Technology, Indore, India
[3]Deptartment of Biological sciences, SHIATS University, Allahabad-211007 (U.P.), India

ABSTRACT: Pathway analysis has become the most popular choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complications and has increased descriptive power. Curators of the metabolic pathway database work to present this information in an easily understandable pathway-based framework. Curators are required to define pathway limitations and classify pathways within a limts of pathway ontology to maximize the utility of the pathways to both researchers and the pathway prediction software. These apparently effortless tasks pose several challenges. This review describes these challenges as well as the criteria that need to be considered, and the rules that have been developed by these curators as they make decisions regarding the representation and classification of metabolic pathway information in different pathway analysis tools.

Keywords: metabolic pathway; metabolic database, Database comparison.

Introduction

Despite the use of advanced technology, analysis of high-throughput data predominantly enlists the differentially expressed genes/proteins. This enlisting of expressed genes is extremely useful in identifying genes that may have significant roles in a biological phenomenon or phenotype. Thus the list has often been unable to predict the mechanistic logics into the underlying biology of the environment being observed for many investigators. In this way, the advent of high-throughput profiling technologies confronts a new test that of extracting meaning from a wide list of differentially expressed genes and proteins.

One way to address this challenge has been to simplify analysis by grouping long lists of individual genes into smaller groups of related genes or proteins. This approach deduces the complexity of analysis. Biologists have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved, as well as how and where gene products interact with each other. Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons.

1. Coupling thousands of genes, proteins and or other biological molecules by the pathways they

- *Kalyani Dhusia*, PH-08765268628 E-mail: kalyanidhusia.bhu@gmail.com*
- *Dr. Ahsan Rizvi, PH-91+7389760144. E-mail: ahsanrizvi@iiti.ac.in*
- *Dr. (Prof.)Pramod W. Ramteke,PH- +91-9415124985E-mail: pwramteke@gmail.com*

Corresponding Author (*)

are involved in reduces the complexity to just several hundred pathways for the experiment.

2. Identifying active pathways that differs between two conditions can have more explainatory power than a simple list of different genes or proteins.

The aim of thid review would be to, firstly describe the existing knowledge base–driven pathway analysis methods, and secondly discuss limitations of each class of methods.

The term ''pathway analysis'' has been used in very broad contexts (**Green, 2006)**. It has been applied to the analysis of Gene Ontology (GO) terms, physical interaction networks (e.g., protein–protein interactions), kinetic simulation of pathways, steady-state pathway analysis (e.g., flux-balance analysis), and in the inference of pathways from expression and sequence data. However, the definition of a ''pathway'' in some of these uses may be incorrect. According to **Merriam Webster,** Pathway is defined as "the sequence of usually enzyme-catalyzed reactions by which one substance is converted into another".

Generations of Pathway analysis Methods

The pathway analysis tools have evolved over decades. They can be broadly categorized under three generations:

1. First generation : the ORAs or over representation analysis approaches
   The immediate need for functional analysis of microarray gene expression data and the emergence of GO gave rise to over-representation analysis (ORA), which statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in its expression.
2. Second generation: FCS or the functional class scoring approaches
   According to functional class scoring (FCS), although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes pathways can also have significant effects. With few exceptions, all FCS methods use a variation of a general framework that consists of the following three steps: first, a gene-level statistic is computed using the molecular measurements from an experiment. This involves computing differential expression of individual genes or proteins.Second, the gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic. The final step in FCS is assessing the statistical significance of the pathway-level statistic.

3. Third generation: Pathway Topology (PT)-Based Approaches PT-based methods are essentially the same as FCS methods in that they perform the same three steps as FCS methods. The key difference between these two is the use of pathway topology to compute gene-level statistics.

Challenges with methodologies

Although widely adopted, the first generation of pathway analysis methods, ORA methods, decouple molecular measurements from functional analysis and assume that genes and pathways are independent of each other. The second-generation FCS methods address these limitations. PT-based methods further improve FCS methods by considering the number and type of interactions between genes, which FCS methods ignore.

There are still outstanding annotation and methodological challenges. First, low resolution knowledge bases, missing condition- and cell-specific information and incomplete annotations restrict development of the next-generation pathway analysis methods. Second, the inability to integrate the dynamic nature of a biological system in analysis limits the utility of existing methods. However, despite these hurdles, as the number and type of functional annotations increase, coupled with technological advances and analysis methods that provide better guidance for strategic planning for subsequent biological experiments, the utility of pathway analysis and confidence in results will likely improve (**Khatri, 2012**).

Let us now see examples of the first generation of pathway analysis tools i.e., Over Representation Analysis approach.

WEGO (Web tool for plotting GO annotations)

GO (gene ontology) consortium provided, unified structured vocabularies and classification which are widely accepted in almost all of the large scale gene annotation projects. As a consequence, many tools have been created for use with the use with the GO ontologies. WEGO (Web gene ontology annotation plot) is a simple and useful tool for comparing visualizing and plotting GO annotation results.

It hasn't been easy to analyze and understand the GO information completely for someone with little computer background. This problem was addressed in 2 ways that are, firstly how to annotate the anonymous sequence with the GO vocabulary and secondly, how to find the deviation or anything noval in the data sets. Many tools and software programs have been developed to tackle the first problem through an automatically or normally curated search for the associations between GO terms and genes (Zehetner,Khan, and Hennig(2003)).

GenMAPP (Gene Microarray Pathway Profiler)

GenMAPP (Gene Microarray Pathway Profiler) is a standalone computer program designed for viewing and analyzing gene expression data in reference to biological pathways. GenMAPP is known to displays gene expression data on pathways by coloring genes based on data and requirement given by the user. GenMAPP also has graphical tools for modifying and constructing pathways. Also, it provides annotation for genes and bridge amongst pathway experts (Eisen, 1998). GenMAPP extends the capabilities of pathway resources by allowing users to design new pathways, to modify pathways for their own use and to use complex strategies for viewing gene expression data on those pathways.GenMAPP represents biological pathways in a special file format called 'MAPPs' (Tomayo, 1999). MAPPs are independent of the gene expression data and can be used to group genes by any organizing principle.

Uses of WEGO: There are two methods to tackle with WEGO. The first is to upload the annotation files (up to 3 files at a time). The input files must be in any one of the four formats input format: WEGO native format, InterProScan raw (our default input format), text and XML output formats. The version of GO archive used for the downstream analysis of the GO annotation results in WEGO should of course be the same as the one used in annotation. Therefore, it is optional in WEGO when uploading the input files. The second way is to simply enter the job ID.A process window showed the job ID after the task is completed. Later the user is redirected to a webpage with a hierarchical GO tree that has all the GO terms within in the uploaded files. The displayed level of Gene Ontology tree and the selected Gene Ontology terms both can be altered by the user.

Researchers can build custom MAPPs with the graphics tools given by the program, providing each gene an identification (ID) from GenBank, SWISSPROT. The gene ID is the link between the gene object on the MAPP, the gene expression data and the annotation for that gene contained in an underlying GenMAPP database (DeRisi, 1997). The annotations, its relative the data and the hyperlinks to the public databases can be accessed by simply one click over the gene.

Uses of geneMAPP: GenMAPP has the flexibility to accept numeric and character data types, calculated values (such as *P* values), data from several experiments and data from both custom-spotted and commercial microarrays. GenMAPP converts the expression data into a data set that can then be viewed on any MAPP with any number of color-coding criteria sets (karp, 2002).

Limitations: Although large number of tools and their widespread use is available, ORA have fewlimitations. For instance, the different statistics used by ORA (e.g., hypergeometric distribution, binomial distribution, chi-square distribution, etc.) are independent of the measured changes. It can be observed from this that these tests consider the number of genes individually and ignore any values associated with it such as probe intensities. By ignoring this information, ORA treat each gene equally. However, the information about the extent of regulation (e.g., fold-changes) can be useful in assigning different weights to input genes, as well as to the pathways they are involved in, which in turn can provide more information than current ORA approaches.

FCS methods address the limitations of ORA. For instance, they do not require an arbitrary threshold for dividing

expression data into significant and non-significant pools. Rather, FCS methods use all available molecular measurements for pathway analysis. Some of the FCS methods using tools are:

SAFE (Significance Analysis of Function and Expression)

SAFEGUI is written in Java in order to provide cross-platform compatibility, and relies upon the Significance Analysis of Function and Expression (SAFE) package (Barry *et al*., 2005) written in R (R Development Core Team, 2006). The release of SAFE 2.0 coincides with the release of SAFEGUI, and adds several new features, including new statistics for differential expression and pathway enrichment, as well as new procedures for error control and resampling. SAFE provides a highly generalized environment for category testing, with a greater variety of options than other resampling category enrichment procedures. When SAFEGUI starts, the user is presented with the main window. The user selects a data file which consists of a row of sample group labels, followed by gene expression measurements. The user selects the appropriate microarray platform and SAFEGUI automatically retrieves the corresponding annotation data from Bioconductor (Gentleman *et al*., 2004).

GeneTrail

GeneTrail is a web-based application that allows the identification of enriched functional categories of protein/gene sets. GeneTrail supports the ORA as well as the GSEA'Gene Set Enrichment Analysis' approach. Also, the implementation of the GSEA analysis involves a novel algorithm that computes the correct p-value instead of calculating it with permutation tests. Since our tool is based on the comprehensive integrative system BN ++ (Kuntzer, 2006), GeneTrail allows the evaluation of a broad range of functional categories. The basic problem with biological data management is the use of appropriate identifier for genes or proteins. Mapping is done for the external identifier to the identifiers being used internally. When NCBI Gene IDs are the internal identifiers and provided data set does not contain NCBI Gene IDs, then GeneTrail needs to convert these IDs.

The user selects a gene-specific 'local' statistic to test for differential expression (e.g. an F -statistic in a one-way ANOVA for a multi-class experimental design). Finally, the user can select from several options to test category enrichment correct for multiple testing of categories/pathways.

SAFEGUI introduces a user-friendly graphical user interface to a powerful statistical package for the analysis of category or pathway significance from microarray data.

Therefore, it is recommended to use the NCBI Gene IDs so as to avoid possible mismatches.

GeneTrail enriches the user with the facility of extracting information from complex proteome data, microarray data or data generated by other high throughput methods with least endeavors.

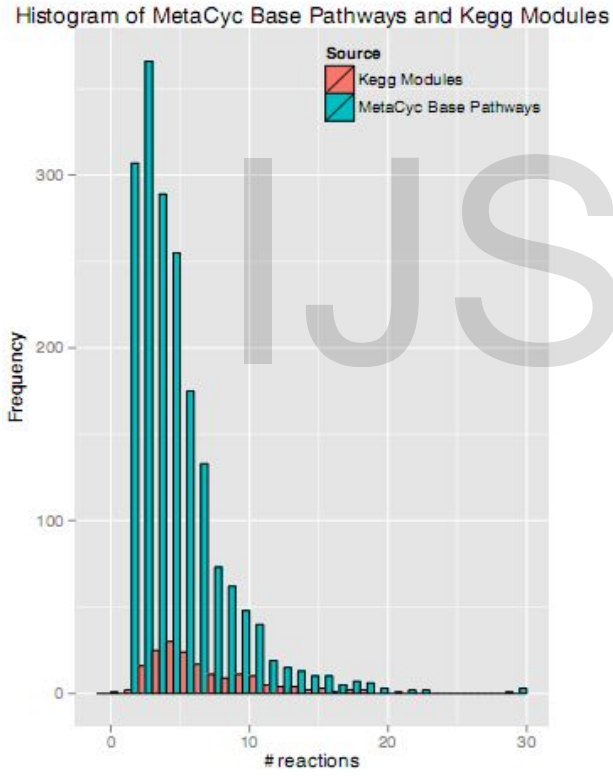
In conclusion, GeneTrail complements the conventional evaluation of experimental data and offers new starting points for further experimental investigation.

Limitations: FCS was an improved methodlogy over ORA as told by **Pavlidis (2004)**, but it also has several limitations. First, similar to ORA, FCS analyzes each pathway independently. This is a limitation because a gene can function in more than one pathway, meaning that pathways can cross and overlap. As a result, if we affect one pathway in any experiment, it is observed that other pathways are also being significantly affected due to the set of overlapping genes. This type of phenomenon is common while using the Gene Ontology terms to define pathways, due to the hierarchical nature of the GO.

Pathway topology (PT)-based methods had been developed to utilize the additional information. PT-based methods are essentially the same as FCS methods in that they perform the same three steps as FCS methods. The key difference between the two is the use of pathway topology to compute gene-level statistics. *Rahnenfuhrer et al.,* proposed ScorePAGE, which computes similarity between each pair of genes in a pathway (**Kanehisa, 2000**). KEGG and MetaCyc are examples of PT based

mathods, lets us make their comparative study.

Relative comparision between MetaCyc compounds and KEGG compounds are made as MetaCyc compounds database links to the corresponding KEGG compounds. The MetaCyc curation staff members add such links during their manual curation. In addition, researchers also submit MetaCyc compounds with chemical structures to the PubChem standardization pipeline in order to compare MetaCyc compound structures with PubChem Compound entries. Periodic processing is done by the same PubChem standardization pipeline for KEGG compounds. A histogram plot of the frequency of MetaCyc base pathway sizes (by reaction count) and KEGG modules sizes (by reaction count) is presented in Figure1, and a histogram plot of the frequency of MetaCyc super pathways and KEGG maps sizes by reaction count is presented in Figure2.



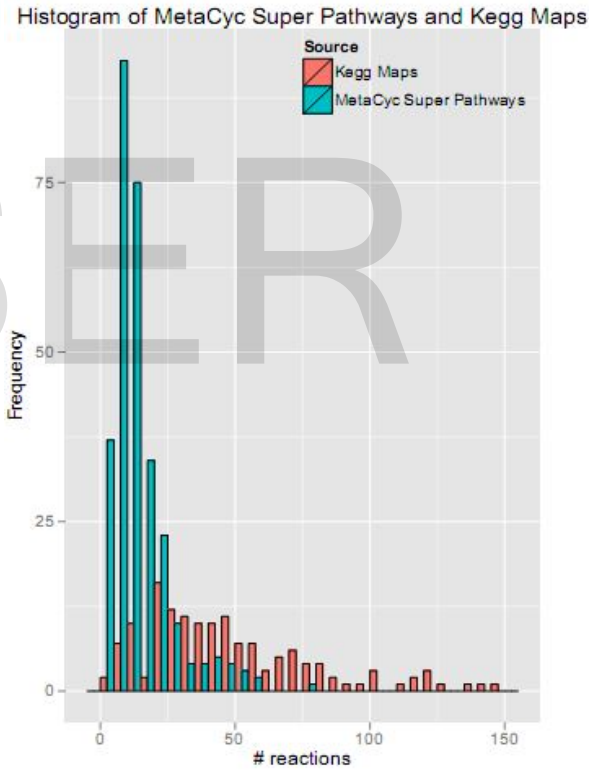


Figure 1: A histogram plot of MetaCyc base pathway and KEGG module size by reaction counts (**TomerAltman, 2013).**

Figure 2: A histogram plot of MetaCyc super pathway and KEGG map size by reaction counts **(TomerAltman, 2013).**

Conclusion:

Consequently the pathway analysis tools have developed and have gone far better over the years. The PT pathways are now in trend and are also quite efficient.

ACKNOWLEDGMENT

We wish to thank the Sam Higginbottom Institute of Agriculture, Technology and Sciences (Formerly Allahabad Agriculture Institute-Deemed University) Deemed University Allahabad, India for providing us the facilities so that we could complete this research work.

References:

[1]. Barry,W.T. *et al* .(2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*,21 , 1943–1949.

[2]. Caspi R, Altman T, Dreher K, Fulcher C. A., Subhraveti P. ,Keseler I ,Kothari A, Krummenacker M, Latendresse M, Mueller L.A. , OngQ ,Paley S, Pujar A, Shearer A G, Travers M, Weerasinghe D, Zhang P, Karp P D :The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nuc Acids R es 2012, 40: D742–D753.

[3]. DeRisi, J.L., Iyer, V.R. & Brown, P.O. *Science* 278, 680–686 (1997)

[4]. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc. Natl Acad. Sci. USA* 95, 14863–14868 (1998)

[5]. Gentleman, R.C. *et al*., (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80

[6]. Green ML, Karp PD (2006) The outcomes of pathway database computations depend on pathway ontology. Nucleic Acids Res 34: 3687–3697.

[7]. Groth D., Lehrach H. and Hennig S. (2004) GOblet: a platform for Gene Ontology annotation of anonymous sequence data. Nucleic Acids Res., 32, W313–W317.

[8]. Hennig,S., Groth,D. and Lehrach,H. (2003) Automated Gene Ontology annotation for anonymous sequence data. Nucleic Acids Res., 31, 3712–3715.

[9]. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

[10]. Karp, P.D. *et al*. *Nucleic Acids Res*. 30, 56–58 (2002).

[11]. Khan, S., Situ, G., Decker, K. and Schmidt, C. J. (2003) Go Figure: automated Gene Ontology annotation. Bioinformatics, 19, 2484–2485.

[12]. Kuntzer J., Blum, T., Gerasch, A., Backes, C., Hildebrandt, A., Kaufmann, M., Kohlbacher, O. and Lenhof H.-P. ,(2006) BN ++ –A biological information system. J. Integr. Bioinformatics 3 , 34.

[13]. Martin, D.M., Berriman, M. and Barton, G.J.(2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes, BMC Bioinformatics, 5, 178.

[14]. Pavlidis P, Qin J, Arango V, Mann J, Sibille E (2004) Using the Gene Ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res 29: 1213–1222.

[15]. Omayo, P. *et al*. *Proc. National Academy Sciences USA* 96, 2907–2912 (1999)

[16]. Tomer Altman, Michael Travers, Anamika Kothari, Ron Caspiand Peter D Karp, A systematic comparison of the MetaCyc and KEGG pathway databases, *BMC Bioinformatics* 2013,14:112

[17]. Zehetner G. (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res., 31, 3799–3803.