Network-Based Approaches for Pathway Level Analysis

Tin Nguyen,[1] Cristina Mitrea,[2] and Sorin Draghici[2,3]

[1]Department of Computer Science and Engineering, University of Nevada, Reno, Nevada
[2]Department of Computer Science, Wayne State University, Detroit, Michigan
[3]Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan

Identification of impacted pathways is an important problem because it allows us to gain insights into the underlying biology beyond the detection of differentially expressed genes. In the past decade, a plethora of methods have been developed for this purpose. The last generation of pathway analysis methods are designed to take into account various aspects of pathway topology in order to increase the accuracy of the findings. Here, we cover 34 such topology-based pathway analysis methods published in the past 13 years. We compare these methods on categories related to implementation, availability, input format, graph models, and statistical approaches used to compute pathway level statistics and statistical significance. We also discuss a number of critical challenges that need to be addressed, arising both in methodology and pathway representation, including inconsistent terminology, data format, lack of meaningful benchmarks, and, more importantly, a systematic bias that is present in most existing methods. © 2018 by John Wiley & Sons, Inc.

Keywords: systems biology • pathway • topology • gene network • survey • pathway analysis

How to cite this article:
Nguyen, T., Mitrea, C., & Draghici, S. (2018). Network-based approaches for pathway level analysis. *Current Protocols in Bioinformatics*, *61*, 8.25.1–8.25.24. doi: 10.1002/cpbi.42

INTRODUCTION

With rapid advances in high-throughput technologies, various kinds of genomic data have become prevalent in most of biomedical research. Advanced techniques in sequencing (e.g., RNA-Seq, miRNA-Seq, DNA-Seq) and microarray assays (e.g., gene expression, methylation) have transformed biological research by enabling comprehensive monitoring of biological systems. Vast amounts of data of all types have accumulated in many public repositories, such as Gene Expression Omnibus (GEO; Barrett et al., 2013; Edgar, Domrachev, & Lash, 2002), Array Express (Brazma et al., 2003; Rustici et al., 2013), The Cancer Genome Atlas (*https://cancergenome.nih.gov/*), and cBioPortal (Cerami et al., 2012; Gao et al., 2013). However, there is a large gap between the ease of data collection and our ability to extract knowledge from these data. Contributing to this gap is the fact that living organisms are complex systems whose emerging phenotypes are the results of multiple complex interactions taking place on various metabolic and signaling pathways.

Regardless of the technology being used, a typical comparative analysis (e.g., disease versus control, treated versus not treated, drug A versus drug B, etc.) often yields a set of genes that are differentially expressed (DE) between the two phenotypes. Even though these lists of DE genes are important in identifying the genes that may be involved in biological changes, they fail to reveal the underlying mechanisms. In order to translate these lists of DE genes into a better understanding of biological phenomena, researchers have developed a variety of knowledge bases that map genes to functional modules. Depending on the amount of information that one wishes to include, these modules can be described as

Current Protocols in Bioinformatics 8.25.1–8.25.24, March 2018
Published online March 2018 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpbi.42

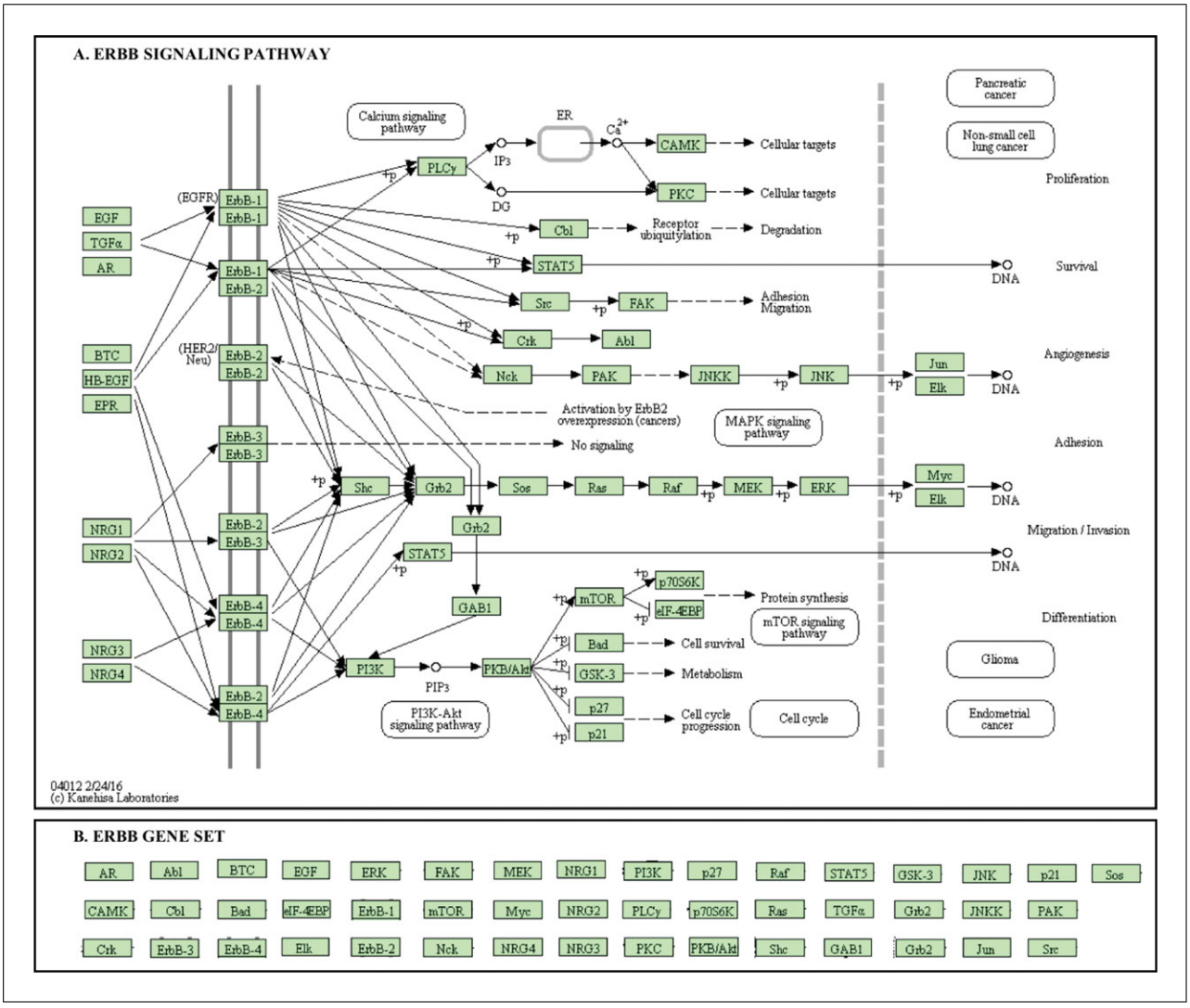

Figure 8.25.1 Pathways are more than gene sets. Panel **A** shows the graphical representation of *ERBB signaling pathway* from KEGG database while panel **B** shows the set of genes on the pathway. The graph in panel A contains important information regarding gene product (protein) localization, gene, protein, or metabolite interactions and the types of these interactions (activation, repression, etc.), the direction of the signal propagation, etc. Over-representation analysis (ORA) and functional class scoring (FCS) approaches are unable to exploit this information. As such, for the same molecular measurement, these approaches would yield exactly the same significance value for this pathway even if the graph were to be completely redesigned by future discovery. In contrast, network-based approaches are able to take into account the topological order of the genes and there interactions.

simple gene sets based on a function, process or component (e.g., the Molecular Signatures Database MSigDB; Liberzon et al., 2011), organized in a hierarchical structure that contains information about the relationship between the various modules, as found in the Gene Ontology (Ashburner et al., 2000), or organized into pathways that describe in detail all known interactions between the various genes that are involved in a certain phenomenon. Biological processes in which genes are known to interact with each other are accumulated in public databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017; Kanehisa & Goto, 2000), Reactome (Croft et al., 2014), and Biocarta (*http://www.biocarta.com*).

Concurrently, statistical methods have been developed to identify the functional modules or pathways that are impacted from the differential expression evidence. They allow us to gain insights into the functional mechanisms of cells beyond the detection of differential expressed genes. The earliest approaches use Over-Representation Analysis (ORA; Drăghici, Khatri, Martins, Ostermeier, & Krawetz, 2003; Tavazoie, Hughes, Campbell, Cho, & Church, 1999) to identify gene sets that have more DE genes than expected

Table 8.25.1 Pathway Analysis Tools

Method	Availability[a]	HIPAA[b]	License[c]	Code[d]	Year	Reference
Signaling pathway analysis using gene expression						
MetaCore[e]	Web (*https://www.genego.com/metacore.php*)	No	Thomson Reuters	Java	2004	N/A
Pathway-Express	Standalone (Bioconductor), Web (*http: //vortex.cs.wayne.edu/*) superseded by the ROntoTools	No	Free[f]	Java, R	2005	(Drăghici et al., 2007, Khatri et al., 2007)
PathOlogist	Standalone (*ftp://ftp1.nci.nih.gov/pub/pathologist/*)	No	Free	MATLAB	2007	(Efroni, Schaefer, & Buetow, 2007; Greenblum, Efroni, Schaefer, & Buetow, 2011)
iPathway Guide[e]	Web (*https://www.advaitabio.com/products.html*)	Yes	Advaita Corp.	Java, R	2009	N/A
SPIA	Standalone (Bioconductor)	No	GPL (≥2)	R	2009	(Tarca et al., 2009)
NetGSA	Standalone (*https://www.biostat.washington.edu/~ashojaie/software/*	No	GPL-2	R	2009	(Shojaie & Michailidis, 2009; Shojaie & Michailidis, 2010)
PWEA	Standalone (*https://zlab.bu.edu/PWEA/*)	No	Free[f]	C++	2010	(Hung et al., 2010)
TopoGSA	Web (*https://www.infobiotics.net/topogsa*)	No	Free[f]	PHP, R	2010	(Glaab, Baudot, Krasnogor, & Valencia, 2010)
Topology GSA	Standalone (CRAN)	No	AGPL-3	R	2010	(Massa, Chiogna, & Romualdi, 2010)
DEGraph	Standalone (Bioconductor)	No	GPL-3	R	2010	(Jacob, Neuvial, & Dudoit, 2010)
GGEA	Standalone (Bioconductor)	No	Artistic-2.0	R	2011	(Geistlinger, Csaba, Küffner, Mulder, & Zimmer, 2011)
BPA	Standalone (*https://bumil.boun.edu.tr/bpa*)	No	Free[f]	MATLAB	2011	(Isci, Ozturk, Jones, & Otu, 2011)
GANPA	Standalone (CRAN)	No	GPL-2	R	2011	(Fang, Tian, & Ji, 2011)
ROnto Tools[e]	Standalone (Bioconductor)	No	CC BY-NC-ND 4	R	2012	(Voichiţa et al., 2012)
BAPA-IGGFD	No implementation available	No	N/A	R	2012	(Zhao et al., 2012)

continued

Table 8.25.1 Pathway Analysis Tools, *continued*

Method	Availability[a]	HIPAA[b]	License[c]	Code[d]	Year	Reference
CePa	Standalone (CRAN), Web (*https://mcube.nju.edu.cn/cgi-bin/cepa/main.pl*)	No	GPL (≥2)	R	2012	(Gu, Liu, Cao, Zhang, & Wang, 2012)
THINK-Back-DS	Standalone, Web (*https://eecs.umich.edu/db/think/software.html*)	No	Free[f]	Java	2012	(Farfán, Ma, Sartor, Michailidis, & Jagadish, 2012)
TBScore	No implementation available	No	N/A	N/A	2012	(Ibrahim, Jassim, Cawthorne, & Langlands, 2012)
ACST	Standalone (available as article supplemental)	No	N/A	R	2012	(Mieczkowski, Swiatek-Machado, & Kaminska, 2012)
EnrichNet	Web (*https://www.enrichnet.org/*)	No	free**	PHP	2012	(Glaab, Baudot, Krasnogor, Schneider, & Valencia, 2012)
clipper	Standalone (*https://romualdi.bio.unipd.it/software*)	No	AGPL-3	R	2013	(Martini, Sales, Massa, Chiogna, & Romualdi, 2013)
DEAP	Standalone (available as article supplemental)	No	GNU Lesser GPL	Python	2013	(Haynes, Higdon, Stanberry, Collins, & Kolker, 2013)
DRAGEN	Standalone (*https://bioinfo.au.tsinghua.edu.cn/dragen/*)	No	N/A	C++	2014	(Ma, Jiang, & Jiang, 2014)
ToPASeq[g]	Standalone (Bioconductor)	No	AGPL-3	R	2016	(Ihnatova & Budinska, 2015)
pDis	Standalone (Bioconductor)	No	Free[f]	R	2016	(Ansari, Voichiţa, Donato, Tagett, & Drăghici, 2017)
SPATIAL	No implementation available	No	N/A	N/A	2016	(Bokanizad, Tagett, Ansari, Helmi, & Drăghici, 2016)
BLMA[h]	Standalone (Bioconductor)	No	GPL (≥2)	R	2017	(Nguyen, Tagett, Donato, Mitrea, & Draghici, 2016; also see Internet Resources)

continued

Table 8.25.1 Pathway Analysis Tools, *continued*

Method	Availability[a]	HIPAA[b]	License[c]	Code[d]	Year	Reference
Signaling pathway analysis using multiple types of data						
PARADIGM	Standalone (*https://sbenz.github.com/Paradigm*)	No	UCSC-CGB, free[f]	C	2010	(Vaske et al., 2010)
micro Graphite	Standalone (*https://romualdi.bio.unipd.it/software*)	No	AGPL-3	R	2014	(Calura et al., 2014)
mirIntegrator	Standalone (Bioconductor)	No	GPL≥3	R	2016	(Diaz et al., 2016; Nguyen et al., 2016)
Metabolic pathway analysis						
ScorePAGE	No implementation available	No	N/A	N/A	2004	(Rahnenführer et al., 2004)
TAPPA	No implementation available	No	N/A	N/A	2007	(Gao & Wang, 2007)
MetPA	Web (*https://metpa.metabolomics.ca*)	No	GPL (≥2)	PHP, R	2010	(Xia & Wishart, 2010)
Metabo Analyst	Web (*https://metpa.metabolomics.ca*)	No	GPL (≥2)	R	2011	(Xia & Wishart, 2011; Xia, Sinelnikov, Han, & Wishart, 2015)

[a]**Availability** is a criterion that describes the implementation of the method as standalone or Web-based.
[b]**HIPAA** provides information about HIPAA compliance.
[c]**License** provides information about the type of the software license. GPL is an abbreviation for the GNU General Public License; AGPL is an abbreviation for the GNU Affero General Public License; CC BY-NC-ND 4 is an abbreviation of Creative Commons Attribution–NonCommercial-NoDerivatives 4.0 International Public License.
[d]**Code** shows the programming language used for the method implementation.
[e]Commercial method.
[f]Free for academic and non-commercial use; UCSC-CGB is the University of California Santa Cruz Cancer Genome Browser.
[g]ToPASeq provides an R package that runs TopologyGSA, DEGraph, clipper, SPIA, TBScore, PWEA, TAPPA.
[h]BLMA provides an R package for bi-level meta-analysis that runs SPIA, ORA, GSA, and PADOG using one or multiple expression datasets.

by chance. The drawbacks of this type of approach include: (i) it only considers the number of DE genes and completely ignores the magnitude of the actual expression changes, resulting in information loss; (ii) it assumes that genes are independent, which they are not (since the pathways are graphs that describe precisely how these genes influence each other); and (iii) it ignores the interactions between various genes and modules. Functional Class Scoring (FCS) approaches, such as Gene Set Enrichment Analysis (GSEA; Subramanian et al., 2005) and Gene Set Analysis (GSA; Efron & Tibshirani, 2007), have been developed to address some of the issues raised by ORA approaches. The main improvement of FCS is the observation that small but coordinated changes in expression of functionally related genes can have significant impact on pathways.

ORA and FCS approaches are often referred as *gene set enrichment* methods. Comprehensive lists of gene set analysis approaches, as well as comparisons between them, can be found in well-developed surveys (Emmert-Streib & Glazko, 2011; Kelder, Conklin, Evelo, & Pico, 2010; Khatri, Sirota, & Butte, 2012; Misman et al., 2009). While useful for the purpose for which they have been developed—to analyze sets of genes—these methods completely ignore the topology and interaction between genes. Topology-based approaches, which fully exploit all the knowledge about how genes interact as described by pathways, have been developed more recently. The first such techniques were ScorePAGE (Rahnenführer, Domingues, Maydt, & Lengauer, 2004) for metabolic pathways and Impact Analysis (Drăghici et al., 2007; Tarca et al., 2009) for signaling

pathways. Figure 8.25.1 shows an example pathway named *ERBB signaling pathway*, in which the nodes represent genes and compounds while the edges represent the known interaction between the compounds. Gene set analysis approaches are not able to account for the topological order of genes, nor are able to explain the signal propagation and mechanisms of the pathway. These approaches would yield exactly the same significance value for this pathway even if the graph were to be completely redesigned by future discovery. In contrast, network-based pathway analysis can distinguish between this pathway and any other pathway with the same proportion of DE genes.

Here we provide a survey of 34 network-based methods developed for pathway analysis. Our survey of commercial tools for pathway analysis found iPathwayGuide (Advaita Corporation, *https://www.advaitabio.com*) and MetaCore (Thomson Reuters, *https://www.thomsonreuters.com*) to be topology based. Other commercial tools, such as Ingenuity Pathway Analysis (*https://www.ingenuity.com*) or Genomatix (*https://www.genomatix.de*), do not use pathway topology information and thus are not included in this review. The existence of only two commercial tools is more evidence of the challenge faced by the developers of such methods due to lack of standards.

In this document, we categorize and compare the methods based on the following criteria: the type of input the method accepts, graph model, and the statistical approaches used to evaluate gene and pathway changes. Under the heading "Experiment Input and Pathway Databases" below we describe the types of input the surveyed methods use. Under the heading "Graph Models" we provide details regarding the graph models used for the biological networks. Under "Pathway Scoring Strategies" we discuss the statistical methods used by the surveyed methods to assess gene and pathway changes between two phenotypes. Under "Challenges in Pathway Analysis" we discuss a number of outstanding challenges that needed to be addressed in order to improve the reliability, as well as the relevance, of the next-generation pathway analysis approaches. We discuss the key elements and limitations of the surveyed methods without going into details of each technique. For a more detailed review and technical descriptions of network-based pathway analysis methods, please refer to Mitrea et al. (2013) or referenced manuscripts (Table 8.25.1).

NETWORK-BASED PATHWAY ANALYSIS

The term *pathway analysis* is used in a very broad context in the literature, including biological network construction and inference. Here we focus on methods that are able to exploit biological knowledge in public repositories, rather than on network inference approaches that attempt to infer or reconstruct pathways from molecular measurements. Table 8.25.1 shows the list of 34 pathway analysis approaches, together with their availability and licensing. In this section we start by describing the availability of each method before discussing the input format, graph model, and underlying statistical approaches.

Software Availability and Implementation

We often think that the main strength of an approach lies in its novelty and algorithm efficiency. However, the implementation and availability of a tool have become increasingly important for several reasons. First, software availability and version control are crucial for reproducing the experimental results that were used to assess the performance of the approach (Sandve, Nekrutenko, Taylor, & Hovig, 2013). For this reason, many journals request authors to make their software and data available before accepting methods articles. Second, if a software application is not ready-to-run, it is very unlikely that the intended audience (mostly life scientists) will invest the time to understand and implement complex algorithms. Practicality, user-friendliness, output format, and type of interface are all to be considered. Depending on the desired availability and intended audience, a software package may be implemented as standalone or Web based. Among the 34 approaches, there are 30 that are available either as a standalone software package or Web service.

Typically, standalone tools need to be installed on local machines or servers, which often requires some administrative skills. Most standalone tools depend on full or partial copies of public pathway databases, stored locally, and need to be updated periodically. Advantages of standalone tools include: (i) instant availability that does not require Internet access, and (ii) the security and privacy of the experimental data. Web-based tools, on the

other hand, run the analyses on a remote server providing computational power and a graphical interface. The major advantage of Web-based tools is that they are user-friendly and do not require a separate local installation. From the accessibility perspective, Web-based tools have the advantage of being available from any location as long as there is an Internet connection and a browser available. Also, the update is almost transparent to the client. This makes the user's task easy and enables collaboration, since users all over the world can utilize the same method without the burden of installing it or keeping it up-to-date. There are methods that provide both Web-based and standalone implementations.

HIPAA compliance may also be a factor in certain applications that involve data coming from patients or data linked to other clinical data or clinical records. Currently, the iPathwayGuide is the only tool available that can do topological pathway analysis and is HIPAA compliant.

The programming language and style used for implementation also play an important role in the acceptance of a method. Software tools that are neatly implemented and packaged are more appealing compared to those that do not have ready-to-use implementations. Many of the methods are implemented in the R programming language and are available as software packages from Bioconductor, CRAN, or the author's Web site. Their popularity among biologists and bioinformaticians is due to the fact that many bioinformatics-dedicated packages are available in R. In addition, the rigorous review procedure provided by Bioconductor and CRAN makes the software packages more standardized and reliable.

Experiment Input and Pathway Databases

A pathway analysis approach typically requires two types of input: experiment data and known biological networks. Experiment data is usually collected from high-throughput experiments that compare a condition phenotype to a control phenotype. A condition phenotype can be a disease, a drug treatment, or the knock-out (KO) of a gene, while a control phenotype can be a healthy state, a different disease or drug treatment, or wild-type (non-KO) samples. Experimental data can be obtained from multiple technologies that produce different types of data: gene expression, protein abundance, metabolite concentration, miRNA expression, etc. Biological networks or pathways are often represented in the form of graphs that capture our current knowledge about the interactions of genes, proteins, metabolites, or compounds in an organism. The pathway data is accumulated, updated, and refined by amassing knowledge from scientific literature describing individual interactions or high-throughput experiment results.

Table 8.25.2 shows a summary of input format and mathematical modeling of the surveyed approaches. Most pathway analysis methods analyze data from high-throughput experiments, such as microarrays, next-generation sequencing, or proteomics. They accept either a list of gene IDs or a list of such gene IDs associated with measured changes. These changes could be measured with different technologies and therefore can serve as proxies for different biochemical entities. For instance, one could use gene expression changes measured with microarrays, or protein levels measured with a proteomic approach, etc.

Different analysis methods use different **input formats**. Many methods accept a list of all genes considered in the experiment together with their expression values. Some analysis methods select a subset of genes, considered to be differentially expressed (DE), based on a predefined cut-off. The cut-off is typically applied on fold-change, *p*-value, or both. These methods use the list of DE genes and their corresponding statistics (fold-change, *p*-value) as input. Other methods use only the list of DE genes, without corresponding expression values, because their scoring methods are based only on the relative positions of the genes in the graph.

Methods which use cut-offs are sensitive to the chosen threshold value, because a small change in the cut-off may drastically change the number of selected genes (Nam & Kim, 2008). In addition, they typically use the most significant genes and discard the rest whose weaker but coordinated changes may also have significant impact on pathways. Genes with moderate differential expression may be lost, even though they might be important players in the impacted pathways (Ben-Shaul, Bergman, & Soreq, 2005). Furthermore, the genes included in the set of DE genes can vary dramatically if the selection methods are changed. Hence, the results of pathway analyses based on DE genes may be vastly different depending on both the selection method as well as the threshold value (Pan, Lih, & Cohen, 2005). Furthermore, for the same disease,

Table 8.25.2 A Summary of the Experimental Data Input Format and Biological Network Databases Used by the Surveyed Methods is Presented

Method	Experiment input[a]	Pathway database[b]	Graph model[c]	Pathway scoring
Signaling pathway analysis using gene expression				
MetaCore	DE genes	Proprietary canonical pathway, genome-scale network	Single-type, directed	Hierarchically aggregated
Pathway-Express	DE genes change, or measured genes change[d]	KEGG signaling	Single-type, directed	Hierarchically aggregated
PathOlogist	Measured genes expression	KEGG	Multi-type, directed	Hierarchically aggregated
iPathwayGuide	DE genes change, or measured genes change	KEGG signaling, Reactome, NCI, BioCarta	Single-type, directed	Hierarchically aggregated
SPIA	DE genes change	KEGG signaling	Single-type, directed	Hierarchically aggregated
NetGSA	Measured genes expression	KEGG signaling	Single-type, directed	Multivariate analysis
PWEA	Measured genes expression	YeastNet	Single-type, undirected	Hierarchically aggregated
TopoGSA	DE genes	PPI network, KEGG	Single-type, undirected	Hierarchically aggregated
TopologyGSA	Measured genes expression	NCI-PID	Single-type, undirected	Multivariate analysis
DEGraph	Measured genes expression	KEGG	Single-type, undirected	Multivariate analysis
GGEA	Measured genes expression	KEGG	Single-type, directed	Aggregate fuzzy similarity
BPA	Measured genes expression—with cut-off	NCI-PID	Single-type, DAG	Bayesian network
GANPA	DE genes change, or measured genes expression	PPI network, KEGG, Reactome, NCI-PID, HumanCyc	Single-type, undirected	Hierarchically aggregated
ROntoTools	DE genes change, or measured gene expression	KEGG signaling	Single-type, directed	Hierarchically aggregated
BAPA-IGGFD	Measured genes expression - with cut-off	Literature-based interaction database; KEGG, WikiPathways; Reactome; MSigDB; GO BP; PANTHER; constructed gene association network from PPIs; co-annotation in GO Biological Process (BP); and co-expression in microarray data	Single-type, DAG	Bayesian network

continued

Table 8.25.2 A Summary of the Experimental Data Input Format and Biological Network Databases Used by the Surveyed Methods is Presented, *continued*

Method	Experiment input[a]	Pathway database[b]	Graph model[c]	Pathway scoring
CePa	DE genes expression, or measured genes expression	NCI-PID	Single-type, directed	Hierarchically aggregated
THINK-Back-DS	DE genes change, measured genes expression	KEGG, PANTHER, BioCarta, Reactome, GenMAPP	Single-type, directed	Hierarchically aggregated
TBScore	DE genes change	KEGG signaling	Single-type, directed	Hierarchically aggregated
ACST	Measured genes expression	KEGG signaling	Single-type, directed	Hierarchically aggregated
EnrichNet	DE genes list	PPI network, KEGG, BioCarta, WikiPathways, Reactome, NCI-PID, InterPro, GO with STRING 9.0	Single-type, undirected	Hierarchically aggregated
clipper	Measured genes expression	BioCarta, KEGG, NCI-PID, Reactome	Single-type, directed	Multivariate analysis
DEAP	Measured genes expression	KEGG, Reactome	Single-type, directed	Hierarchically aggregated
DRAGEN	Measured genes expression	RegulonDB, M3D, HTRIdb, ENCODE, MSigDB	Single-type, directed	Linear regression
ToPASeq[e]	Measured genes expression	KEGG	Single-type, directed	Hierarchical & multivariate
pDis	DE genes change, or measured genes change[d]	KEGG signaling	Single-type, directed	Hierarchically aggregated
SPATIAL	DE genes change	KEGG signaling	Single-type, directed	Hierarchically aggregated
BLMA[f]	Measured genes expression	KEGG signaling	Single-type, directed	Hierarchically aggregated
Signaling pathway analysis using multiple types of data				
PARADIGM	Measured genes expression, copy number, and proteins levels	Constructed PPI networks from MIPS, DIP, BIND, HPRD, IntAct, and BioGRID	Multi-type, directed	Hierarchically aggregated
microGraphite	Measured gene expression, and miRNA expression	BioCarta, KEGG, NCI-PID, Reactome	Single-type, directed	Multivariate analysis
mirIntegrator	DE genes change, and DE miRNA change	KEGG signaling, miRTarBase	Single-type, directed	Hierarchically aggregated
Metabolic pathway analysis				
ScorePAGE	Measured genes expression	KEGG metabolic	Single-type, undirected	Hierarchically aggregated

continued

Table 8.25.2 A Summary of the Experimental Data Input Format and Biological Network Databases Used by the Surveyed Methods is Presented, *continued*

Method	Experiment input[a]	Pathway database[b]	Graph model[c]	Pathway scoring
TAPPA	Measured genes expression	KEGG metabolic	Single-type, undirected	Hierarchically aggregated
MetPA	DE metabolites change	KEGG metabolic	Single-type, directed	Hierarchically aggregated
MetaboAnalyst	DE genes change and DE metabolites change	KEGG metabolic	Single-type, directed	Hierarchically aggregated

[a]**Experiment input** shows the experiment data input format for each method. "DE" means differentially expressed. "Change" means fold-change value or *t*-statistics when comparing gene/metabolites values between two phenotypes. "Measured" means the list of all the genes/metabolites measured in the experiment; "List" represents a list of genes/metabolites identifiers (e.g., symbols). "With cut-off" show methods that take as input the list of all measured genes and in the analysis they mark the DE genes.
[b]**Pathway database** shows the name of the knowledge source for biological interactions.
[c]**Graph model** is a characteristic that shows if the graph has one or multiple types of nodes as well as if directed or undirected.
[d]The package ROntoTools (Bioconductor) allows for analysis using expression values of all genes.
[e]ToPASeq provides an R package that runs TopologyGSA, DEGraph, clipper, SPIA, TBScore, PWEA, TAPPA.
[f]BLMA provides an R package for bi-level meta-analysis that runs SPIA, ORA, GSA, and PADOG using one or multiple expression datasets.

independent studies or measurements often produce different sets of differentially expressed (DE) genes (Ein-Dor, Kela, Getz, Givol, & Domany, 2005; Ein-Dor, Zuk, & Domany, 2006; Tan et al., 2003). This makes approaches that use DE genes as input appear even more unreliable.

Usually, pathways are sets of genes and/or gene products that interact with each other in a coordinated way to accomplish a given biological function. A typical **signaling pathway** (in KEGG for instance) uses nodes to represent genes or gene products and edges to represent signals, such as activation or repression, that go from one gene to another. A typical **metabolic pathway** uses nodes to represent biochemical compounds and edges to represent reactions that transform one or more compound(s) into one or more other compounds. These reactions are usually carried out or controlled by enzymes, which are in turn coded by genes. Hence, in a metabolic pathway, genes or gene products are associated with edges rather than nodes, as in a signaling pathway. The immediate consequence of this difference is that many techniques cannot be applied directly on all available pathways. This is why the analysis of metabolic pathways is still generally and arguably underdeveloped (only 128 Google Scholar citations to date for the original ScorePage paper; Rahnenführer et al., 2004) and there are not many other methods available for metabolic pathway analysis. However, the topology-based analysis of signaling pathways has been very successful (over 1200 citations to date for the two impact analysis papers mentioned above; Drăghici et al., 2007; Tarca et al., 2009) and over 30 other methods have been developed since.

There are other types of biological networks that incorporate genome-wide interactions between genes or proteins such as **protein-protein interaction (PPI)** networks. These networks are not restricted to specific biological functions. The main caveat related to PPI data is that most such data are obtained from a bait-prey laboratory assays, rather than from in vivo or in vitro studies. The fact that two proteins stick to each other in an assay performed in an artificial environment can be misleading, since the two proteins may never be present at the same time in the same tissue or the same part of the cell.

Publicly available curated pathway databases used by methods listed in Table 8.25.2 are KEGG (Ogata et al., 1999), NCI-PID (Schaefer et al., 2009), BioCarta (*https://www.biocarta.com*), WikiPathways (Pico et al., 2008), PANTHER (Mi et al., 2005), and Reactome (Joshi-Tope et al., 2005). These knowledge bases are built by manually curating experiments performed in different cell types under different conditions. These curated knowledgebases are more reliable than protein interaction networks but do not include all known genes and their interactions. As an example, despite being continuously updated, KEGG includes only about 5,000 human genes in signaling pathways while the number of protein-coding

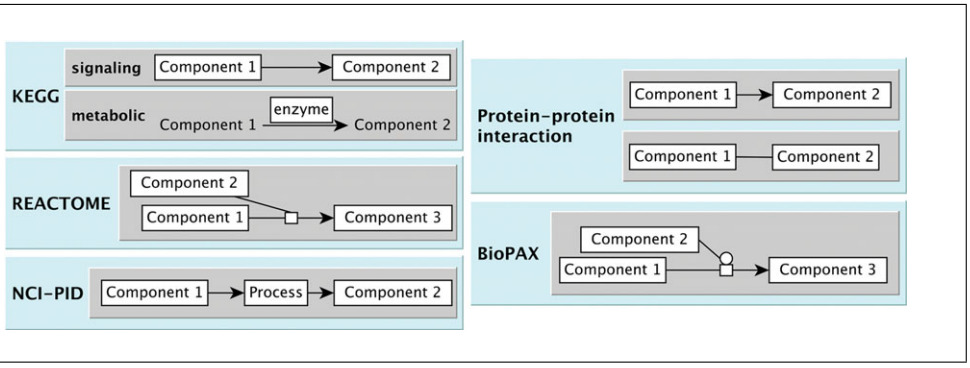

Figure 8.25.2 Five biological network graph models used by public databases are displayed. The **KEGG** database contains both signaling and metabolic networks. **Signaling** networks have genes/gene products as nodes and regulatory signals as edges. Types of regulatory signals include activation, inhibition, phosphorylation, and many others. **Metabolic** networks have biochemical compounds as nodes and chemical reactions as edges. Enzymes (specialized proteins/gene products) catalyze biochemical reactions; therefore, genes are linked to the edges in these networks. The **Reactome** database is a collection of biochemical reactions that are grouped in functionally related sets to form a pathway. There are two types of nodes: biochemical compounds and reactions. Reaction nodes link biochemical compounds as reactants and products. **NCI-PID** signaling networks also have two types of nodes: component nodes and process nodes. Component nodes are usually biomolecular components. Process nodes are usually biochemical reactions or biological processes. In these networks, process nodes link two or more component nodes through directed edges. Process nodes are assigned one of the following states: positive or negative regulation, or involved in. **Protein-protein interaction (PPI)** networks have proteins as nodes and physical binding as edges or interactions. Two-hybrid assays are typically used to determine protein-protein interactions. PPIs can be directed in the bait-prey orientation when the bait-prey relation is considered (top), or undirected when is not (bottom). **The Biological Pathway Exchange (BioPAX)** format has physical entities as nodes and conversions as edges. The advantage of this representation is that it is generic and provides a lot of flexibility to accommodate various types of interactions. In addition, it provides a machine-readable standard that can be used for all databases to provide their data in a unified format. Network nodes can be genes, gene products, complexes, or non-coding RNA. Network edges can be assembly of a complex, disassembly of a complex, or biochemical reactions, among others.

genes is estimated to be between 19,000 and 20,000 (Ezkurdia et al., 2014).

The implementation of analysis methods constrains the software to accept a specific input pathway data format, while the underlying graph models in the methods are independent of the input format. Regardless of the pathway format, this information must be parsed into a computer-readable graph data structure before being processed. The implementation may incorporate a parser, or this may be up to the user. For instance, SPIA accepts any signaling pathway or network if it can be transformed into an adjacency matrix representing a directed graph where all nodes are components and all edges are interactions. NetGSA is similarly flexible with regard to signaling and metabolic pathways. SPIA provides KEGG signaling pathways as a set of pre-parsed adjacency matrices. The methods described in this chapter may be restricted to only one pathway database, or may accept several.

To the best of our knowledge, there have been no comprehensive approaches to compare the pros and cons of existing knowledge bases. It is completely up to researchers to choose the tools that are able to work with the database they trust. Intuitively, methods that are able to exploit the complementary information available from different databases have an edge over methods that work with one single database.

Graph Models

Pathway analysis approaches use two major graph models to represent biological networks obtained from knowledge bases: (i) single-type and (ii) multiple-type. The first model allows only one type of node, i.e., a gene or protein, with edges representing molecular interactions between the nodes (KEGG signaling in Fig. 8.25.2). Models that contain directed graphs are more suitable for analyses that include signal propagation of gene perturbation. The second graph model allows multiple type of nodes, such as components and interactions (NCI-PID in Fig. 8.25.2). Multi-type graph models are more complex than

single-type, but they are expected to capture more pathway characteristics. For example, single-type models are limited when trying to describe "all" and "any" relations between multiple components that are involved in the same interaction. Bipartite graphs, which contain two types of nodes and allow connection only between nodes of different types, are a particular case of multi-type graph models.

The majority of analysis methods surveyed here use a single-type graph model. Some apply the analysis on a directed or un-directed single-type network built using the input pathway, while others transform the pathways into graphs with specific characteristics. An example of the later is TopologyGSA, which transforms the directed input pathway into an undirected decomposable graph, which has the advantage of being easily broken down into separate modules (Lauritzen, 1996). In this method, decomposable graphs are used to find "important" submodules—those that drive the changes across the whole pathway. For each pathway, TopologyGSA creates an undirected moral graph from the underlying directed acyclic graph (DAG) by connecting the parents of each child and removing the edge direction. [The moral graph of a DAG is the undirected graph created by adding an (undirected) edge between all parents of the same node (sometimes called marrying), and then replacing all directed edges by undirected edges. The name stems from the fact that, in a moral graph, two nodes that have a common child are required to be married by sharing an edge.] The moral graph is then used to test the hypothesis that the underlying network is changed significantly between the two phenotypes. If the the research hypothesis is rejected, a decomposable/triangulated graph is generated from the moral graph by adding new edges. This graph is broken into the maximal possible submodules and the hypothesis is retested on each of them.

PathOlogist and PARADIGM are the two surveyed methods that use multi-type graph models. PathOlogist uses a bipartite graph model with component and interaction nodes. PARADIGM, conceptually motivated by the central dogma of molecular biology, takes a pathway graph as input and converts it into a more detailed graph, where each component node is replaced by several more specific nodes: biological entity nodes, interaction nodes, and nodes containing observed experiment data. The observed experiment nodes could in principle contain gene-expression and copy-number information. Biological entity nodes are DNA, mRNA, protein, and active protein. The interaction nodes are transcription, translation, or protein activation, among others. Biological entity and interaction node values are derived from these data and specify the probability of the node being active. These are the hidden states of the model. From the mathematical model perspective, models that allow multiple types of nodes—component nodes and interaction nodes—are more flexible and are able to model both AND and OR gates, which are very common when describing cellular processes.

Pathway Scoring Strategies

The goal of the scoring method is to compute a score for each pathway based on the expression change and the graph model, resulting in a ranked list of pathways or sub-pathways. There are a variety of approaches to quantify the changes in a pathway. Some of the analysis methods use a hierarchically aggregated scoring algorithm, where on the first level a score is calculated and assigned to each node or pair of nodes (component and/or interaction). On the second level, these scores are aggregated to compute the score of the pathway. On the last level, the statistical significance of the pathway score is assessed using univariate hypothesis testing. Another approach assigns a random variable to each node, and a multivariate probability distribution is calculated for each pathway. The output score can be calculated in two ways. One way is to use multivariate hypothesis testing to assess the statistical significance of changes in the pathway distribution between the two phenotypes. The other way is to estimate the distribution parameters based on the Bayesian network model and use this distribution to compute a probabilistic score to measure the changes. In this section, we provide details regarding the scoring algorithms of the surveyed methods. See Figure 8.25.3 for scoring algorithms categories.

Hierarchically aggregated scoring

The workflow of the hierarchically aggregated scoring strategy has three levels: node statistic computation, pathway statistic computation and the evaluation of the significance for the pathway statistic.

Node level scoring

Most of the surveyed methods incorporate pathway topology information in the node scores. The node level scoring can be divided into four categories: (i) graph measures (centrality), (ii) similarity measures,

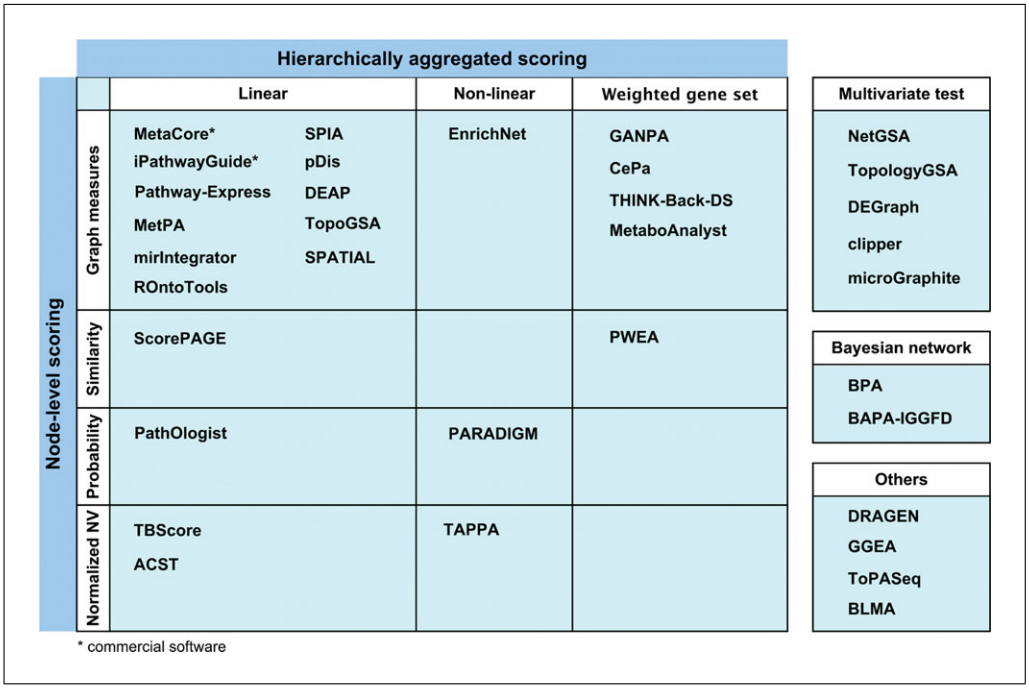

Figure 8.25.3 A summary of the statistical models is presented for the surveyed methods. Most methods use the *hierarchically aggregated scoring* strategy, in which the score is computed at the node level before being aggregated at the pathway level for significance assessment. In the left panel, the rows show a node-level statistical model while the columns show the aggregating strategies (linear, non-linear, or weighted gene set). Approaches in *multivariate analysis* and *Bayesian network* categories use multivariate modeling and Bayesian network, respectively, to find pathways that are most likely to be impacted. Both BLMA and ToPASeq provide R packages that run multiple algorithms. DRAGEN follows a linear regression strategy while GGEA applies fuzzy modeling to rank the pathways.

(iii) probabilistic graphical models, and (iv) normalized node value (NNV). Approaches in the first category use centrality measures or a variation of these measures to score nodes in a given pathway. Centrality measures represent the importance of a node relative to all other nodes in a network. There are several centrality measures that can be applied to networks of genes and their interactions: degree centrality, closeness, between-ness, and eigenvector centrality. Degree centrality accounts for the number of directed edges that enter and leave each node. Closeness sums the shortest distance from each node to all other nodes in the network. Node between-ness measures the importance of a node according to the number of shortest paths that pass through it. Eigenvector centrality uses the network adjacency matrix of a graph to determine a dominant eigenvector; each element of this vector is a score for the corresponding node. Thus, each score is influenced by the scores of neighboring nodes. In the case of directed graphs, a node that has many downstream genes has more influence and receives a higher score. Methods in this category include MetaCore, SPIA, iPathwayGuide, MetPA, SPATIAL, TopoGSA, DEAP, pDis, mirIntegrator, Pathway-Express, and BLMA.

Approaches in the second category use similarity measures in their node level scoring. Similarity measures estimate the co-expression, behavioral similarity, or co-regulation of pairs of components. Their values can be correlation coefficients, covariances, or dot products of the gene expression profile across time or samples. In these methods, the pathways with clusters of highly correlated genes are considered more significant. At the node level, a score is assigned to each pair of nodes in the network which is the ratio of one similarity measure over the shortest path distance between these nodes. Thus, the topology information is captured in the node score by incorporating the shortest path distance of the pair. Methods in this category include ScorePAGE and PWEA.

Approaches in the third category incorporate the topology in the node level scoring using a probabilistic graphical model. In this

model, nodes are random variables, and edges define the conditional dependency of the nodes they link. For example, PARADIGM takes observed experimental data and calculates scores for all component nodes, in both observed and hidden states, from the detailed network created by the method based on the input pathway. For each node score, a positive or negative value denotes how likely it is for the node to be active or inactive, respectively. The scores are calculated to maximize the probability of the observed values. A *p*-value is associated with each score of each sample such that each node can be tagged as significantly active, significantly inactive, or not significant. For each network, a matrix of *p*-values is output in which columns are samples and rows are component nodes. Methods in this category include PARADIGM and PathOlogist.

Approaches in the fourth category simply compute the score for each node using the information obtained from the experiment input. For example, TAPPA calculates the score of each node as the square root of the normalized log gene expressions (node value) while ACST and TBScore calculate the node level score using a sign statistic and log fold-change.

Pathway level scoring

There are three different ways to compute the pathway score: (i) linear, (ii) non-linear, and iii) weighted gene set. Most methods aggregate node level statistics to pathway level statistics using linear functions such as averaging or summation. For example, iPathwayGuide computes the scores of pathway as the sum of all genes while TBScore weights the pathway DE genes based on their log fold change and the number of distinct DE genes directly downstream of them, using a depth-first search algorithm.

Approaches in the second group (non-linear) use a non-linear function to compute the pathway scores. For example, TAPPA computes the pathway score for each sample as a weighted sum of the product of all node pair scores in the pathway. The weight coefficient is 0 when there is no edge between a pair. For any connected node pair the weight is a sign function, which represents joint up- or down-regulation of the pair. Another example is EnrichNet, in which pathway scores measure the difference of the node score distribution for a pathway and a background network/gene set which consists of all pathways. At the node level, the distance of all DE genes to the pathway is measured and summarized as a distance distribution. The method assumes that the most relevant pathway is the one with the greatest difference between the pathway node score distribution and the background score distribution. The difference between the two distributions is measured by the weighted averaging of the difference between the two discretized and normalized distributions. The averaging method down-weights the higher distances and emphasizes the lower distance nodes.

Methods in the third group (weighted gene set) design scoring techniques that incorporate existing gene set analysis methods, such as GSEA (Subramanian et al., 2005), GSA (Efron & Tibshirani, 2007), or LRPath (Sartor, Leikauf, & Medvedovic, 2009). Pathway-level scores can be calculated using node scores that represent the topology characteristic of the pathway as weight adjustments to a gene set analysis method. PWEA, GANPA, THINK-Back-DS, and CePa use this approach, and we refer to them as weighted gene set analysis methods.

Pathway significance assessment

Pathway scores are intended to provide information regarding the amount of change incurred in the pathway between two phenotypes. However, the amount of change is not meaningful by itself, since any amount of change can take place just by chance. An assessment of the *significance* of the measured changes is thus required.

TopoGSA, MetPA, and EnrichNet output scores without any significance assessment, leaving it up to the user to interpret the results. This is problematic because users do not have any instrument to distinguish between changes due to noise or random causes and meaningful changes that are unlikely to occur just by chance and that therefore are possibly related to the phenotype. The rest of the analysis methods perform a hypothesis test for each pathway. The null hypothesis is that the value of the observed statistic is due to random noise or chance alone. The research hypothesis is that the observed values are substantial enough that they are potentially related to the phenotype. A *p*-value for calculated score is then computed, and a user-defined threshold on the *p*-value is used to decide whether the the null hypothesis can be rejected or not for each pathway. Finally, a correction for multiple comparisons should be performed.

Typically, pathway analysis methods compute one score per pathway. The distribution of this score under the null hypothesis can be constructed and compared to the observed score. However, there are often too few samples to

calculate this distribution, so it is assumed that the distribution is known. For example, in MetaCore and many other techniques, when the pathway score is the number of DE nodes that fall on the pathway, the distribution is assumed to be hypergeometric. However, the hypergeometric distribution assumes that the variables (genes in this case) are independent, which is incorrect, as witnessed by the fact that the pathway graph structure itself is designed to reflect the specific ways in which the genes influence each other. Another approach to identify the distribution is to use statistical techniques such as the bootstrap (Efron, 1979). Bootstrapping can be done either at the sample level, by permuting the sample labels, or at gene-set level, by permuting the the values assigned to the genes in the set.

Multivariate analysis and Bayesian network

Multivariate analysis methods mostly use multivariate probability distributions to compute pathway-level statistics and these can be grouped into two subcategories. Methods in the first category use multivariate hypothesis testing, while methods in the second category are based on Bayesian network.

NetGSA, TopologyGSA, DEGraph, clipper, and microGraphite are methods based on multivariate hypothesis testing. These analysis methods assume the vectors of gene expression values in each (sub)pathway are random vectors with multivariate normal distributions. The network topology information is stored in the covariance matrix of the corresponding distribution. For a network, if the two distributions of the gene expression vectors corresponding to the two phenotypes are significantly different, the network is assumed to be significantly impacted when comparing the two phenotypes. The significance assessment is done by a multivariate hypothesis test. The definition of the null hypothesis for the statistical tests and the techniques to calculate the parameters of the distributions are the main differences between these three analysis methods.

BPA and BAPA-IGGFD are two methods based on Bayesian networks. In a Bayesian network, which is a special case of probabilistic graphical models, a random variable is assigned to each node of a directed acyclic (DAG) graph. The edges in the graph represent the conditional probabilities between nodes, so that the children are independent from each other and the rest of the graph when conditioned on the parents. In BPA, the value of the Bayesian random variable assigned to each node captures the state of a gene (DE or not). In contrast, in BAPA-IGGFD each random variable assigned to an edge is the probability that up or down regulation of the genes at both ends of an interaction are concordant with the type of interaction which can be activation or inhibition. In both BPA and BAPA-IGGFD, each random variable is assumed to follow a binomial distribution whose probability of success follows a beta distribution. However, these two methods use different approaches in representing the multivariate distribution of the corresponding random vector. BPA assumes that the random vector has a multinomial distribution, which is the generalization of the binomial distribution. In this case, the vector of the success probability follows the Dirichlet distribution, which is the multivariate extension of the beta distribution. In contrast, BAPA-IGGFD assumes the random variables are independent, therefore the multivariate distributions are calculated by multiplying the distributions of the random variables in the vector. It is worth mentioning that the assumption of independence in BAPA-IGGFD is contradicted by evidence, specifically in the case of edges that share nodes.

Other approaches

The four methods DRAGEN, GGEA, ToPASeq, and BLMA follow strategies that are very different from those of the other 30 methods. As such, BLMA implements a bi-level meta-analysis approach that can be applied in conjunction with any of the four statistical approaches SPIA (Tarca et al., 2009), GSA (Efron & Tibshirani, 2007), ORA (Tavazoie et al., 1999), or PADOG (Tarca, Drăghici, Bhatti, & Romero, 2012). The package allows users to perform pathway analysis with one dataset or with multiple datasets. Similarly, ToPASeq provides an R package that runs TopologyGSA, DEGraph, clipper, SPIA, TBScore, PWEA, and TAPPA. This method models the input (biological networks and experiment data) in such a way that it can be used for any of the seven different analyses.

DRAGEN is an analysis method that scores the interactions rather then the genes and uses a regression model to detect differential regulation. DRAGEN fits each edge of a pathway into two linear models (for case and control) and then computes a p-value that represents the difference between the two models. For each pathway, a summary statistic is computed by combining the p-values of the edges using a weighted Fisher's method (Fisher, 1925).

GGEA, on the other hand, uses Petri Net (Murata, 1989) to model the pathway. The summary statistic, named consistency, is computed from the fuzzy similarity between the observed gene expression and Petri net with fuzzy logic (PNFL; Küffner, Petri, Windhager, & Zimmer, 2010). For both DRAGEN and GGEA, the *p*-value of the pathway is calculated by comparing the observed summary statistic against the null distribution that is constructed by permutation.

CHALLENGES IN PATHWAY ANALYSIS

Pathway analysis has become the first choice for gaining insights into the underlying biology of a phenotype due its explanatory power. However, there are outstanding annotation and methodological limitations that have not been addressed (Kelder, Conklin, Evelo, & Pico, 2010; Khatri et al., 2012). There are three main limitations of current knowledge bases. First, existing knowledge bases are unable to keep up with the information available in data obtained from recent technologies. For example, RNA-Seq data allows us to identify transcripts that are active under certain conditions. Alternatively spliced transcripts, even if they originate from the same gene, may have distinct or even opposite functions (Wang et al., 2008). However, most knowledge bases provide pathway annotation only at the gene level. Second, there is a lack of condition and cell-specific information, i.e., information about cell type, conditions, and time points. Finally, current pathway annotations are neither complete nor perfectly accurate (Khatri et al., 2012; Rhee, Wood, Dolinski, & Drăghici, 2008). For example, the number of genes in KEGG have remained around 5,000 despite being updated continuously in the past 10 years. while there are approximately 19,000 human genes annotated with at least one GO term. The number of protein-coding genes is estimated to be between 19,000 and 20,000 (Ezkurdia et al., 2014), most of which are included in DNA microarray assays, such as Affymetrix HG U133 plus 2.0.

Another challenge is the oversimplification that characterizes many of the models provided by pathway databases. In principle, each type of tissue might have different mechanisms, so generic, organism-level pathways present a somewhat simplistic description of the phenomena. Furthermore, signaling and metabolic processes can also be different from one condition to another, or even from one patient to another. Understanding the specific pathways that are impacted in a given phenotype or sub-group of patients should be another goal for the next generation of pathway analysis tools. See Khatri et al. (2012) for a more detailed discussion of annotation limitations of existing knowledge bases.

Here, we focus on challenges of pathway analysis from computational perspectives. We demonstrate that there is a systematic bias in pathway analysis (Nguyen, Mitrea, Tagett, & Drăghici, 2017). This leads to the unreliability of most if not all pathway analysis approaches. We also discuss the lack of benchmark datasets or pipelines to assess the performance of existing approaches.

Systematic Bias of Pathway Analysis Methods

Pathway analysis approaches often rely on hypothesis testing to identify the pathways that are impacted under the effects of different diseases. Null distributions are used to model populations so that statistical tests can determine whether an observation is unlikely to occur by chance. In principle, the *p*-values produced by a sound statistical test must be uniformly distributed under the null hypothesis (Barton, Crozier, Lillycrop, Godfrey, & Inskip, 2013; Bland, 2013; Fodor, Tickle, & Richardson, 2007; Storey & Tibshirani, 2003). For example, the *p*-values that result from comparing two groups using a *t*-test should be distributed uniformly if the data are normally distributed (Bland, 2013). When the assumptions of statistical models do not hold, the resulting *p*-values are not uniformly distributed under the null hypothesis. This makes classical methods, such as *t*-test inaccurate since gene expression values do not necessary follow their assumptions. Here, we also show that the problem is extended to pathway analysis, as the pathway *p*-values obtained from statistical approaches are not uniformly distributed under the null hypothesis. This might lead to severe bias towards well-studied diseases, such as cancer, and thus make the results unreliable (Nguyen et al., 2017).

Consider three pathway analysis methods that represent three different classes of methods for pathway analysis: Gene Set Analysis (GSA; Efron & Tibshirani, 2007) is a Functional Class Scoring method (Efron & Tibshirani, 2007; Mootha et al., 2003; Subramanian et al., 2005; Tarca et al., 2012), Down-weighting of Overlapping Genes (PADOG; Tarca et al., 2012) is an enrichment method (Beißbarth & Speed,

2004; Drăghici et al., 2003; Khatri, Drăghici, Ostermeier, & Krawetz, 2002), and Signaling Pathway Impact Analysis (SPIA; Tarca et al., 2009) is a topology-aware method (Drăghici et al., 2007; Tarca et al., 2009). To simulate the null distribution, we download and process the data from nine public datasets: GSE14924 CD4, GSE14924 CD8, GSE17054, GSE12662, GSE57194, GSE33223, GSE42140, GSE8023, and GSE15061. Using 140 control samples from the nine datasets, we simulate 40,000 datasets as follows. We randomly label 70 samples as *control* samples and the remaining 70 samples as *disease* samples. We repeat this procedure 10,000 times to generate different groups of 70 control and 70 disease samples. To make the simulation more general, we also create 10,000 datasets consisting of 10 control and 10 disease samples, 10,000 datasets consisting of 10 control and 20 disease samples, and 10,000 datasets consisting of 20 control and 10 disease samples. We then calculate the *p*-values of the KEGG human signaling pathways using each of the three methods.

The effect of combining control (i.e., healthy) samples from different experiments is to uniformly distribute all sources of bias among the random groups of samples. If we compare groups of control samples based on experiments, there could be true differences due to batch effects. By pooling them together, we form a population which is considered the reference population. This approach is similar to selecting from a large group of people that may contain different sub-groups (e.g., different ethnicities, gender, race, life style, or living conditions). When we randomly select samples (for the two random groups to be compared) from the reference population, we expect all bias (e.g., ethnic subgroups) to be represented equally in both random groups, and therefore, we should see no difference between these random groups, no matter how many distinct ethnic subgroups were present in the population at large. Therefore, the *p*-values of a test for difference between the two randomly selected groups should be equally probable between zero and one.

Figure 8.25.4 displays the empirical null distributions of *p*-values using PADOG, GSA, and SPIA. The horizontal axes represent p-values while the vertical axes represent *p*-value densities. Green panels (A0-A6) show *p*-value distributions from PADOG, while blue (B0-B6) and purple (C0-C6) panels show *p*-value distributions from GSA and SPIA, respectively. For each method, the larger panel (A0, B0, C0) shows the cumulative *p*-values from all KEGG signaling pathways. The small panels, six per method, display extreme examples of nonuniform *p*-value distributions for specific pathways. For each method, we show three distributions severely biased towards zero (e.g., A1-A3), and three distributions severely biased towards one (e.g., A4-A6).

These results show that, contrary to generally accepted beliefs, the *p*-values are not uniformly distributed for the three methods considered. Therefore, one should expect a very strong and systematic bias in identifying significant pathways for each of these methods. Pathways that have *p*-values biased towards zero will often be falsely identified as significant (false positives). Likewise, pathways that have *p*-values biased towards one are likely to rarely meet the significance requirements, even when they are truly implicated in the given phenotype (false negatives). Systematic bias, due to nonuniformity of *p*-value distributions, results in failure of the statistical methods to correctly identify the biological pathways implicated in the condition, and also leads to inconsistent and incorrect results.

Querying Data from Knowledge Bases

Independent research groups have tried different strategies to model complex biomolecular phenomena. These independent efforts have led to variation among pathway databases, complicating the task of developing pathway analysis methods. Depending on the database, there may be differences in information sources, experiment interpretation, models of molecular interactions, or boundaries of the pathways. Therefore, it is possible that pathways with the same designation and aiming to describe the same phenomena may have different topologies in different databases. As an example, one could compare the insulin signaling pathways of KEGG and BioCarta. BioCarta includes fewer nodes and emphasizes the effect of insulin on transcription, while KEGG includes transcription regulation as well as apoptosis and other biological processes. Differences in graph models for molecular interactions are particularly apparent when comparing the signaling pathways in KEGG and NCI-PID. While KEGG represents the interaction information using the directed edges themselves, NCI-PID introduces "process nodes" to model interactions (see Fig. 8.25.2). Developers are facing the challenge of modifying methods to accept

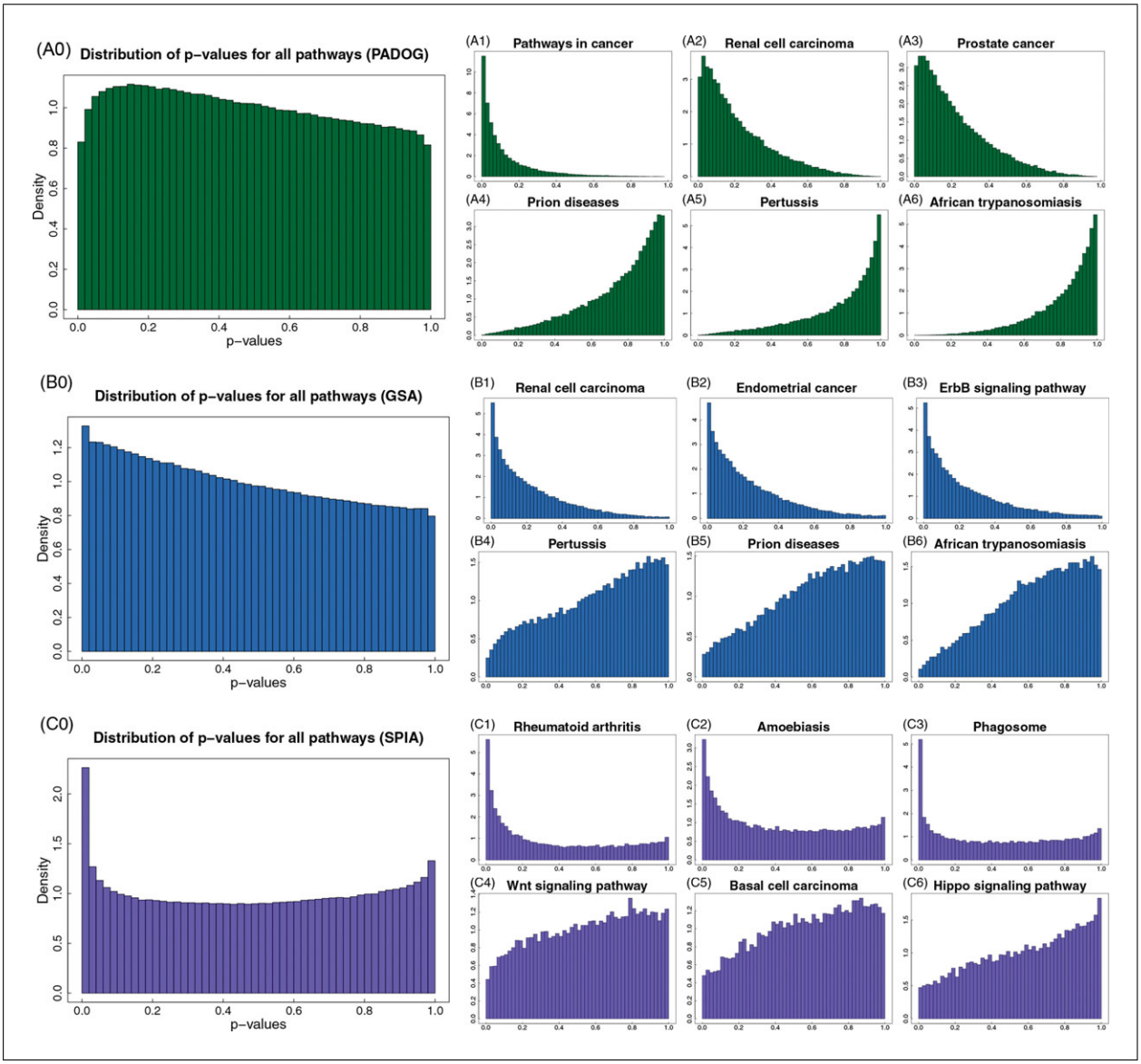

Figure 8.25.4 The empirical distributions of *p*-values using: Down-weighting of Overlapping Genes (PADOG; top), Gene Set Analysis (GSA; middle), and Signaling Pathway Impact Analysis (SPIA; bottom). The distributions are generated by re-sampling from 140 control samples obtained from 9 AML datasets. The horizontal axes display the *p*-values, while the vertical axes display the *p*-value densities. Panels A0-A6 (green) show the distributions of *p*-values from PADOG; panels B0-B6 (blue) show the distribution of *p*-values from GSA; panels C0-C6 (purple) show the distribution of *p*-values from SPIA. The large panels on the left, A0, B0, and C0, display the distributions of *p*-values cumulated from all KEGG signaling pathways. The smaller panels on the right display the *p*-value distributions of selected individual pathways, which are extreme cases. For each method, the upper three distributions, for example A1-A3, are biased towards zero and the lower three distributions, for example A4-A6, are biased towards one. Since none of these *p*-value distributions are uniform, there will be systematic bias in identifying significant pathways using any one of the methods. Pathways that have *p*-values biased towards zero will often be falsely identified as significant (false positives). Likewise, pathways that have *p*-values biased towards one are more likely to be among false negative results even if they may be implicated in the given phenotype.

novel pathway databases or modifying the actual pathway graphs to conform to the method.

Pathway databases not only differ in the way that interactions are modeled, but their data are provided in different formats as well (Chuang, Hofree, & Ideker, 2010). Common formats are Pathway Interaction Database eXtensible Markup Language (**PID XML**), KEGG Markup Language (**KGML**), Biological Pathway Exchange (**BioPAX**) Level 2 and Level 3, System Biology Markup Language (**SBML**), and the Biological Connection Markup Language (**BCML**; Beltrame et al., 2011). The NCI provides a unified assembly of BioCarta and Reactome, as well as their in-house "NCI-Nature curated

pathways," in NCI-PID format (Schaefer et al., 2009). In order to unify pathway databases, pathway information should be provided in a commonly accepted format.

Another challenge is that the same biological pathways are represented differently from one pathway database to another. None of the tools is compatible with all database formats, requiring either modification of pathway input or alteration of the underlying algorithm in order to accommodate the differences. As an example, a study by Vaske et al. (2010) attempts to compare SPIA (Tarca et al., 2009) with their tool PARADIGM by re-implementing SPIA in C and forcing its application on NCI-PID pathways. Implementation errors are present in the C version of SPIA, invalidating the comparison. A solution to overcome this challenge could be the development of a unified globally accepted pathway format. Another possible solution is to build conversion software tools that can translate between pathway formats. Some attempts exist to use BIO-PAX (Demir et al., 2010) as the lingua franca for this domain.

Missing Benchmark Datasets and Comparison

Newly developed approaches are typically assessed by simulated data or by well-studied biological datasets (Bayerlova et al., 2015; Varadan, Mittal, Vaske, & Benz, 2012). The advantage of using simulation is that the ground truth is known and can be used to compare the sensitivity and specificity of different methods. However, simulation is often biased and does not fully reflect the complexity of living organisms. On the other hand, when using real biological data, the biology is never fully known. In addition, many papers presenting new pathway analysis methods include results obtained on only a couple of datasets, and researchers are often influenced by the observer-expectancy effect (Sackett, 1979). Thus, such results are not objective, and many times they cannot be reproduced.

A better evaluation approach has been proposed by Tarca et al. (2013) using 42 real datasets. This approach uses a target pathway which is the pathway describing the condition under study available. For instance, in an experiment with colon cancer versus healthy, the target pathway would be the colon cancer pathway. The datasets are chosen so that there is a target pathway associated with each of the datasets. The datasets are also all public, so other methods can be compared on the very same data in a reproducible and objective way. The lower the rank and the p-value of the target pathway in the method output, the better the method. This approach has several important advantages including the fact that it is reproducible and completely objective, and relies on more than just a couple of data sets that are assessed by the authors using the literature. This approach was also used by Bayerlova et al. (2015), using a different set of 36 real datasets, as well by other, more recent papers.

However, in spite of its advantages and great superiority compared to the usual method of only analyzing a couple of data sets, even this benchmarking has important limitations. First, not all conditions have a namesake pathway in existing databases or described in the literature. Second, complex diseases are often associated with not only one target pathway, but with many biological processes. By its nature, this assessment approach will ignore other pathways and their ranking, even though they may be true positives. More importantly, these approaches fail to take into consideration the systematic bias of pathway analysis approaches. In these review papers, most of the datasets are related to cancer. As such, 28 out of 42 datasets used in Tarca et al. (2013), and 26 out of 36 datasets used in Bayerlova et al. (2015) are cancer datasets. In those cases, methods that are biased towards cancer are very likely to identify cancer pathways, which are also the target pathways, as significant. For this reason, the comparisons obtained from these reviews are likely to be biased and not reliable for assessing the performance of existing approaches.

CONCLUSIONS

Pathway analysis is a core strategy of many basic research, clinical research, and translational medicine programs. Emerging applications range from targeting and modeling disease networks to screening chemical or ligand libraries, to identification of drug target interactions for improved efficacy and safety. The integration of molecular interaction information into pathway analysis represents a major advance in the development of mathematical techniques aimed at the evaluation of systems perturbations in biological entities.

This unit discussed and categorized 34 existing network-based pathway analysis approaches from different perspectives, including experiment input, graphical representation of pathways in knowledge bases, and statistical approaches to assess pathway significance.

Despite being widely used, employing DE genes as input makes the software sensitive to cutoff parameters. Regarding the graph models, approaches using multiple types of nodes and bipartite graphs are more flexible and are able to model both AND and OR gates, which are very common when describing cellular processes. In addition, tools that are able to work with multiple knowledge bases are expected to perform better due to their complementary and independent information that cannot be obtained from individual databases. We also pointed out that there has been no reliable benchmark to assess and rank existing approaches. Some initial efforts to assess pathway analysis methods in an objective and reproducible way do exist, but they still fail to take into account the statistical bias of existing approaches.

Despite tremendous efforts in the field, there are outstanding challenges that need to be addressed. First, current pathway databases are unable to provide transcript-level activity or information related to new types of data, such as SNP, mutation, or methylation. Second, incomplete annotation and the lack of condition- and cell-specific information hinder the accuracy of downstream pathway analysis. Third, variation among pathway databases and the lack of a standard format in which the pathway data is provided pose a real challenge for implementation. Developers are facing the challenge of modifying methods to accept novel pathway databases or modifying the actual pathway graphs to conform with their method. Fourth, there is a systematic bias due to the fact that certain conditions, such as cancer, are much more studied than others. Using control data obtained from nine mRNA datasets, we showed that the *p*-values obtained by three pathway analysis approaches that represent three mainstream strategies in pathway analysis (FCS, enrichment, and network-based) are not uniformly distributed under the null. Pathways that have *p*-values biased towards zero will often be falsely identified as significant (false positives). Likewise, pathways that have *p*-values biased towards one are likely to rarely meet the significance requirements, even when they are truly implicated in the given phenotype (false negatives). Systematic bias, due to non-uniformity of *p*-value distributions, results in failure of the statistical methods to correctly identify the biological pathways implicated in the condition, and also leads to inconsistent and incorrect results.

Finally, there is a lack of benchmarks to assess the performance of each mathematical approach, as well as to validate existing knowledge bases and their graphical representation of pathways. There have been some efforts to provide benchmarks for this purpose but such methods are not fully reliable yet.

CONFLICT-OF-INTEREST STATEMENT

Sorin Draghici is the founder and CEO of Advaita Corporation, a Wayne State University spin-off that commercializes iPathwayGuide, one of the pathway analysis tools mentioned in this chapter.

ACKNOWLEDGMENTS

This work has been partially supported by the following grants: National Institutes of Health (R01 DK089167, R42 GM087013), National Science Foundation (DBI-0965741), and the Robert J. Sokol, M.D. Endowed Chair in Systems Biology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

LITERATURE CITED

Ansari, S., Voichiţa, C., Donato, M., Tagett, R., & Drăghici, S. (2017). A novel pathway analysis approach based on the unexplained disregulation of genes. *Proceedings of the IEEE*, *105*(3), 482–495. doi: 10.1109/JPROC.2016.2531000.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*, 25–29. doi: 10.1038/75556.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Research*, *41*(D1), D991–D995. doi: 10.1093/nar/gks1193.

Barton, S. J., Crozier, S. R., Lillycrop, K. A., Godfrey, K. M., & Inskip, H. M. (2013). Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions. *BMC Genomics*, *14*(1), 161 doi: 10.1186/1471-2164-14-161.

Bayerlova, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., & Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics*, *16*(1), 334. doi: 10.1186/s12859-015-0751-5.

Beißbarth, T., & Speed, T. P. (2004). GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, *20*, 1464–1465. doi: 10.1093/bioinformatics/bth088.

Beltrame, L., Calura, E., Popovici, R. R., Rizzetto, L., Guedez, D. R., Donato, M., ... Cavalieri, D. (2011). The biological connection markup language: A SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics*, *27*(15), 2127–2133. doi: 10.1093/bioinformatics/btr339.

Ben-Shaul, Y., Bergman, H., & Soreq, H. (2005). Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, *21*(7), 1129–1137. doi: 10.1093/bioinformatics/bti149.

Bland, M. (2013). Do baseline *p*-values follow a uniform distribution in randomised trials? *PloS One*, *8*(10), e76010. doi: 10.1371/journal.pone.0076010.

Bokanizad, B., Tagett, R., Ansari, S., Helmi, B. H., & Drăghici, S. (2016). SPATIAL: A system-level PAThway impact AnaLysis approach. *Nucleic Acids Research*, *44*(11), 5034–5044. doi: 10.1093/nar/gkw429.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., ... Sansone, S-A. (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *31*(1), 68–71. doi: 10.1093/nar/gkg091.

Calura, E., Martini, P., Sales, G., Beltrame, L., Chiorino, G., D'Incalci, M., ... Romualdi, C. (2014). Wiring miRNAs to pathways: A topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Research*, *42*(11), e96. doi: 10.1093/nar/gku354.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., ... Larsson, E. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401–404. doi: 10.1158/2159-8290.CD-12-0095.

Chuang, H.-Y., Hofree, M., & Ideker, T. (2010). A decade of systems biology. *Annual Review of Cell and Developmental Biology*, *26*, 721–744. doi: 10.1146/annurev-cellbio-100109-104122.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., ... D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(D1), D472–D477. doi: 10.1093/nar/gkt1102.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., ... Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, *28*(9), 935–942. doi: 10.1038/nbt.1666.

Diaz, D., Donato, M., Nguyen, T., & Draghici, S. (2016). MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume *22*, page 390. NIH Public Access.

Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., & Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, *81*(2), 98–104. doi: 10.1016/S0888-7543(02)00021-6.

Drăghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichiţa, C., ... Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, *17*(10), 1537–1545. doi: 10.1101/gr.6202607.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. doi: 10.1093/nar/30.1.207.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. doi: 10.1214/aos/1176344552.

Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, *1*(1), 107–129. doi: 10.1214/07-AOAS101.

Efroni, S., Schaefer, C. F., & Buetow, K. H. (2007). Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*, *2*(5), e425. doi: 10.1371/journal.pone.0000425.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2005). Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics*, *21*(2), 171–178. doi: 10.1093/bioinformatics/bth469.

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *In Proceedings of the National Academy of Sciences*, *103*(15), 5923–5928. doi: 10.1073/pnas.0601231103.

Emmert-Streib, F., & Glazko, G. V. (2011). Pathway analysis of expression data: Deciphering functional building blocks of complex diseases. *PLoS Computational Biology*, *7*(5), e1002053. doi: 10.1371/journal.pcbi.1002053.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., ... Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, *23*(22), 5866–5878. doi: 10.1093/hmg/ddu309.

Fang, Z., Tian, W., & Ji, H. (2011). A network-based gene-weighting approach for pathway analysis. *Cell Research*, *22*(3), 565–580. doi: 10.1038/cr.2011.149.

Farfán, F., Ma, J., Sartor, M. A., Michailidis, G., & Jagadish, H. V. (2012). THINK Back: Knowledge-based interpretation of high throughput data. *BMC Bioinformatics*, *13*(Suppl 2), S4. doi: 10.1186/1471-2105-13-S2-S4.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fodor, A. A., Tickle, T. L., & Richardson, C. (2007). Towards the uniform distribution of null P values on Affymetrix microarrays. *Genome Biology*, *8*(5), R69. doi: 10.1186/gb-2007-8-5-r69.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, Sur., ... Schultz, N. (2013). Integrative analysis of complex cancer genomics

and clinical profiles using the cBioPortal. *Science Signaling*, *6*(269), pl1. doi: 10.1126/scisignal.2004088.

Gao, S., & Wang, X. (2007). TAPPA: Topological analysis of pathway phenotype association. *Bioinformatics*, *23*(22), 3100–3102. doi: 10.1093/bioinformatics/btm460.

Geistlinger, L., Csaba, G., Küffner, R., Mulder, N., & Zimmer, R. (2011). From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, *27*(13), i366–i373. doi: 10.1093/bioinformatics/btr228.

Glaab, E., Baudot, A., Krasnogor, N., & Valencia, A. (2010). TopoGSA: Network topological gene set analysis. *Bioinformatics*, *26*(9), 1271–1272. doi: 10.1093/bioinformatics/btq131.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. (2012). EnrichNet: Network-based gene set enrichment analysis. *Bioinformatics*, *28*(18), i451–i457. doi: 10.1093/bioinformatics/bts389.

Greenblum, S., Efroni, S., Schaefer, C., & Buetow, K. (2011). The PathOlogist: An automated tool for pathway-centric analysis. *BMC Bioinformatics*, *12*(1), 133. doi: 10.1186/1471-2105-12-133.

Gu, Z., Liu, J., Cao, K., Zhang, J., & Wang, J. (2012). Centrality-based pathway enrichment: A systematic approach for finding significant pathways dominated by key genes. *BMC Systems Biology*, *6*(1), 56. doi: 10.1186/1752-0509-6-56.

Haynes, W. A., Higdon, R., Stanberry, L., Collins, D., & Kolker, E. (2013). Differential expression analysis for pathways. *PLoS Computational Biology*, *9*(3), e1002967. doi: 10.1371/journal.pcbi.1002967.

Hung, J.-H., Whitfield, T. W., Yang, T.-H., Hu, Z., Weng, Z., & DeLisi, C. (2010). Identification of functional modules that correlate with phenotypic difference: The influence of network topology. *Genome Biology*, *11*(2), R23. doi: 10.1186/gb-2010-11-2-r23.

Ibrahim, M. A.-H., Jassim, S., Cawthorne, M. A., & Langlands, K. (2012). A topology-based score for pathway enrichment. *Journal of Computational Biology*, *19*(5), 563–573. doi: 10.1089/cmb.2011.0182.

Ihnatova, I., & Budinska, E. (2015). ToPASeq: An R package for topology-based pathway analysis of microarray and RNA-Seq data. *BMC Bioinformatics*, *16*(1), 350. doi: 10.1186/s12859-015-0763-1.

Isci, S., Ozturk, C., Jones, J., & Otu, H. H. (2011). Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics*, *27*(12), 1667–1674. doi: 10.1093/bioinformatics/btr269.

Jacob, L., Neuvial, P., & Dudoit, S. (2010). Gains in power from structured two-sample tests of means on graphs. *Arxiv preprint arXiv:1009.5173*.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., ... Stein, L. (2005). REACTOME: A knowledgebase of biological pathways. *Nucleic Acids Research*, *33*(Database issue), D428–D432. doi: 10.1093/nar/gki072.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. doi: 10.1093/nar/28.1.27.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. doi: 10.1093/nar/gkw1092.

Kelder, T., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2010). Finding the right questions: Exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS Biology*, *8*(8), e1000472. doi: 10.1371/journal.pbio.1000472.

Kelder, T., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2010). Finding the right questions: Exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS Biology*, *8*(8), e1000472. doi: 10.1371/journal.pbio.1000472.

Khatri, P., Drăghici, S., Ostermeier, G. C., & Krawetz, S. A. (2002). Profiling gene expression using Onto-Express. *Genomics*, *79*(2), 266–270. doi: 10.1006/geno.2002.6698.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2), e1002375. doi: 10.1371/journal.pcbi.1002375.

Khatri, P., Voichiţa, C., Kattan, K., Ansari, N., Khatri, A., Georgescu, C., ... Drăghici, S. (2007). Onto-Tools: New additions and improvements in 2006. *Nucleic Acids Research*, *35*(Web Server issue), W206–W211. doi: 10.1093/nar/gkm327.

Küffner, R., Petri, T., Windhager, L., & Zimmer, R. (2010). Petri nets with fuzzy logic (PNFL): Reverse engineering and parametrization. *PLoS One*, *5*(9), e12807. doi: 10.1371/journal.pone.0012807.

Lauritzen, S. L. (1996). *Graphical models* (Vol. *17*). Oxford University Press.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. doi: 10.1093/bioinformatics/btr260.

Ma, S., Jiang, T., & Jiang, R. (2014). Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data. *Bioinformatics*, *31*(4), 563–571. doi: 10.1093/bioinformatics/btu672.

Martini, P., Sales, G., Massa, M. S., Chiogna, M., & Romualdi, C. (2013). Along signal paths: An empirical gene set approach exploiting pathway topology. *Nucleic Acids Research*, *41*(1), e19–e19. doi: 10.1093/nar/gks866.

Massa, M. S., Chiogna, M., & Romualdi, C. (2010). Gene set analysis exploiting the topology of a

pathway. *BMC Systems Biology*, *4*(1), 121. doi: https://doi.org/10.1186/1752-0509-4-121.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., ... Thomas, P. D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, *33*(Suppl 1), D284–D288. doi: 10.1093/nar/gki078.

Mieczkowski, J., Swiatek-Machado, K., & Kaminska, B. (2012). Identification of pathway deregulation–gene expression based analysis of consistent signal transduction. *PLoS One*, *7*(7), e41541. doi: 10.1371/journal.pone.0041541.

Misman, M. F., Deris, S., Hashim, S. Z. M., Jumali, R., & Mohamad, M. S. (2009). Pathway-based microarray analysis for defining statistical significant phenotype-related pathways: A review of common approaches. In *Information Management and Engineering, 2009. ICIME'09. International Conference on*, pages 496–500. Piscataway, NJ: IEEE.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., ... Drăghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, *4*, 278. doi: 10.3389/fphys.2013.00278.

Mootha, V. K., Lindgren, C. M., Eriksson, K-F., Subramanian, A., Sihag, S., Lehar, J., ... Groop, L. C. (2003). PGC-11α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, *34*(3), 267–273. doi: 10.1038/ng1180.

Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, *77*(4), 541–580. doi: 10.1109/5.24143.

Nam, D., & Kim, S.-Y. (2008). Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, *9*(3), 189–197. doi: 10.1093/bib/bbn001.

Nguyen, T., Diaz, D., Tagett, R., & Draghici, S. (2016). Overcoming the matched-sample bottleneck: An orthogonal approach to integrate omic data. *Nature Scientific Reports*, *6*, 29251. doi: 10.1038/srep29251.

Nguyen, T., Mitrea, C., Tagett, R., & Drăghici, S. (2017). DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions—applied to biological pathway analysis. *Proceedings of the IEEE*, *105*(3), 496–515. doi: 10.1109/JPROC.2015.2507119.

Nguyen, T., Tagett, R., Donato, M., Mitrea, C., & Draghici, S. (2016). A novel bi-level meta-analysis approach-applied to biological pathway analysis. *Bioinformatics*, *32*(3), 409–416. doi: 10.1093/bioinformatics/btv588.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*(1), 29–34. doi: 10.1093/nar/27.1.29.

Pan, K.-H., Lih, C.-J., & Cohen, S. N. (2005). Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(25), 8961–8965. doi: 10.1073/pnas.0502674102.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., & Evelo, C. (2008). WikiPathways: Pathway editing for the people. *PLoS Biology*, *6*(7), e184. doi: 10.1371/journal.pbio.0060184.

Rahnenführer, J., Domingues, F. S., Maydt, J., & Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, *3*(1). doi: 10.2202/1544-6115.1055.

Rhee, Y. S., Wood, V., Dolinski, K., & Drăghici, S. (2008). Use and misuse of the Gene Ontology annotations. *Nature Reviews Genetics*, *9*(7), 509–515. doi: 10.1038/nrg2363.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., ... Sarkans, U. (2013). ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Research*, *41*(D1), D987–D990. doi: 10.1093/nar/gks1174.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*(1-2), 51–63. doi: 10.1016/0021-9681(79)90012-2.

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, *9*(10), e1003285. doi: 10.1371/journal.pcbi.1003285.

Sartor, M. A., Leikauf, G. D., & Medvedovic, M. (2009). LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, *25*(2), 211–217. doi: 10.1093/bioinformatics/btn592.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., ... Buetow, K. H. (2009). PID: The pathway interaction database. *Nucleic Acids Research*, *37*(Suppl 1), D674–D679. doi: 10.1093/nar/gkn653.

Shojaie, A., & Michailidis, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, *16*(3), 407–426. doi: 10.1089/cmb.2008.0081.

Shojaie, A., & Michailidis, G. (2010). Network enrichment analysis in complex experiments. *Statistical Applications in Genetics and Molecular Biology*, *9*(1). doi: 10.2202/1544-6115.1483.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445. doi: 10.1073/pnas.1530509100.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences of*

the Unites States of America, *102*(43), 15545–15550. doi: 10.1073/pnas.0506580102.

Tan, P. K., Downey, T. J., Spitznagel, E. L. Jr, Xu, P., Fu, D., Dimitrov, D. S., ... Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, *31*(19), 5676–5684. doi: 10.1093/nar/gkg763.

Tarca, A. L., Bhatti, G., & Romero, R. (2013). A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS One*, *8*(11), e79217. doi: 10.1371/journal.pone.0079217.

Tarca, A. L., Drăghici, S., Bhatti, G., & Romero, R. (2012). Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, *13*(1), 136. doi: https://doi.org/10.1186/1471-2105-13-136.

Tarca, A. L., Drăghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., ... Romero, R. (2009). A novel signaling pathway impact analysis (SPIA). *Bioinformatics*, *25*(1), 75–82. doi: 10.1093/bioinformatics/btn577.

Tarca, A. L., Drăghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., ... Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, *25*(1), 75–82. doi: 10.1093/bioinformatics/btn577.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, *22*, 281–285. doi: 10.1038/10343.

Varadan, V., Mittal, P., Vaske, C. J., & Benz, S. C. (2012). The integration of biological pathway knowledge in cancer genomics: A review of existing computational approaches. *Signal Processing Magazine, IEEE*, *29*(1), 35–50. doi: 10.1109/MSP.2011.943037.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., ... Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), i237–i245. doi: 10.1093/bioinformatics/btq182.

Voichiţa, C., Donato, M., & Drăghici, S. (2012). Incorporating gene significance in the impact analysis of signaling pathways. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume *1*, pages 126–131, Boca Raton, FL, USA, 12–15 December 2012. Piscataway, NJ: IEEE.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., ... Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470. doi: 10.1038/nature07509.

Xia, J., & Wishart, D. S. (2010). MetPA: A web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, *26*(18), 2342–2344. doi: 10.1093/bioinformatics/btq418.

Xia, J., & Wishart, D. S. (2011). Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nature Protocols*, *6*(6), 743. doi: 10.1038/nprot.2011.319.

Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Research*, *43*(W1), W251–W257. doi: 10.1093/nar/gkv380.

Zhao, Y., Chen, M.-H., Pei, B., Rowe, D., Shin, D-G., Xie, W., ... Kuo, L. (2012). A Bayesian approach to pathway analysis by integrating gene–gene functional directions and microarray data. *Statistics in Biosciences*, *4*(1), 105–131. doi: 10.1007/s12561-011-9046-1.

Internet Resources

https://www.biocarta.com

BioCarta: Charting Pathways of Life.

https://bioconductor.org/packages/release/bioc/manuals/BLMA/man/BLMA.pdf

BLMA: A package for bi-level meta-analysis. R package version 1