Send Orders for Reprints to reprints@benthamscience.ae

Current Bioinformatics, **2017**, *12*, **000-000**

REVIEW ARTICLE

A Review of Software Tools for Pathway Crosstalk Inference

Julieta S. Dussaut, Rocío L. Cecchini, Cristian A. Gallo, Ignacio Ponzoni and Jessica A. Carballido*

Instituto de Ciencias e Ingeniería de la Computación (ICIC). Universidad Nacional del Sur-CONICET, San Andrés 800 – Campus Palihue, 8000 Bahía Blanca, Argentina

ARTICLE HISTORY

Received: June 15, 2016
Revised: September 6, 2016
Accepted: November 9, 2016

DOI: 10.2174/1574893611666161123123204

Abstract: The inference of different kinds of relations between pathways constitutes a challenging step towards the analysis of biological data. In this regard, this review article aims at outlining several methods that analyze associations between pathways starting from different sources of information, namely the internet, databases, and/or gene expression data. The source used in each case constitutes a first criterion for the classification of the methods. In this sense, some methods are strongly supported by the pathway topology annotations, whereas others can infer relationships extracting free-topology associations. The second criterion for grouping the methods is based on the inference strategies. The advantages and drawbacks of each methodology are presented, as well as a taxonomy tree and summary table as an overview of the discussion.

Keywords: Relations between pathways, topological pathway information, microarray data, pathway enrichment, inference methods.

1. INTRODUCTION

We are living in an era that is in general characterized by a lot of data but little information. That is, there are thousands of experimental numbers, but it is hard to interpret them in order to obtain valuable information. This is strongly evidenced in the area of systems biology, where everyday researchers seek to achieve an understanding of different biological processes at the molecular level, starting from what is sometimes called "big data". Fortunately, improvements and innovations in technology continue to stimulate the quality and types of biological data that can be obtained at the genome level. Thus, a lot of data collected over several years is now presented as annotations and databases. In this context, all this data properly combined and grouped has great potential for enabling novel discoveries which would then, finally and hopefully, lead to advances in biology and medicine [1].

The focus in this article is on outlining the methods that perform a special analysis of genomic data in the research area of bioinformatics. In this sense, the basic motivation consists in understanding biological phenomena that occur at the cellular level, so it is necessary to unravel the complex mechanisms of interaction between different cellular processes. Specifically, the objective is to review various strategies that address this task by identifying different kinds of relations between biological pathways. This involves detecting signaling mechanisms between biological processes from the coordination that takes place between genes underlying the different pathways [2-8].

*Address correspondence to this autor at the Instituto de Ciencias e Ingeniería de la Computación (ICIC). Universidad Nacional del Sur-CONICET, San Andrés 800 – Campus Palihue, 8000 Bahía Blanca, Argentina; E-mail: jac@cs.uns.edu.ar

With that problem in mind, the aim of this work is to present a review of computational techniques for inferring *interactions* between pathways from microarray expression data and also from different databases with curated data. The strategies revised belong to different computational areas, such as evolutionary computation, text mining, and statistical analysis.

Modelling Pathway's Interactions

A pathway consists of a series of actions between molecules in a cell that generate a certain product or change therein. Such a route may trigger the formation of new molecules, proteins for example. Biological pathways can also control genes, or stimulate a cell to move. Several types of biological pathways exist, the most common being involved in metabolism, signal transmission or gene regulation. Currently, it is known that biological pathways are much more complex than was once thought. Most of the routes do not have a start or end point; in fact, many of them have no real limits and often work together to perform certain tasks. Therefore, when multiple biological pathways *interact* with each other they form a biological network which is often called a pathway network [9, 10].

Existing methods presented in the literature use alternative means for inferring pathway networks; some propose to analyze interactions between biological pathways often called crosstalk, while others allude to finding differentially expressed routes based on the data provided. To achieve their results these methods work with data from microarray, or simply use topological information of the biological pathways granted by databases such as KEGG [11, 12]; they can also take advantage of a combination of both approaches. Moreover, in the methods identified in this

1574-8936/17 $58.00+.00

article we include some that use data mining, and others that are based on mathematical and statistical calculations. In the following section of the paper, strategies that aim at building pathway networks will be introduced, classified and analyzed.

2. INFERENCE METHODS

The aim of this section is to present a categorization of the strategies that find different kinds of relationships between pathways, consequently building a pathway network. A first approach that can be used to infer a network of pathways consists in evaluating when two pathways are differentially expressed, thus supposing that they might be related. Methods that detect which pathways are differentially expressed are reviewed in section 2.1. Next, methods that directly search for relations between pathways will be grouped in sections 2.2 and 2.3 according to whether or not they use microarray data.

Therefore, section 2.1 constitutes an overview, where we introduce four methods that do not directly find networks of pathways, but instead analyze pathway differentiation. They are included in the article as they constitute inspiration for some of the other methods presented. Also, even though they do not directly find networks of pathways, they are important since they could be used to infer some relations between pathways that are differentially expressed together.

2.1. Pathways that are Differentially Expressed Together

DNA microarray data are commonly used to provide a comparison of the expression patterns of genes under "control vs. affected" conditions, during the study of a given disease. Since this comparison usually reveals a large number of differentially expressed genes, it is often difficult, if not impossible, to analyze each gene individually. Therefore strategies that aim at identifying sets of biologically relevant genes appear in the literature; these groups of genes are generally called "pathway enrichment" or "gene set enrichment". The detection of differentially expressed genes helps to associate biological phenotypes with their underlying molecular mechanisms, thus providing a vision of their biologic function.

One of the typical applications of microarray technology consists in the identification of differentially expressed genes according to two conditions. The most common statistical approach for this is to quantify the relevance of each gene with a p-value that can be adjusted for multiple comparisons [13]. Then, an appropriate threshold is set and a list of candidate genes is created. This approach is often disapproved of since it ignores biological knowledge about how genes work together. The next three methods start from the premise of improving this drawback and are often used to find a biological explanation concerning the abundant amounts of data of differentially expressed genes.

GSEA

This review starts with the description of GSEA as it is very frequently mentioned in the literature. The algorithm owes its acronym to "Gene Set Enrichment Analysis" [2]. In this case, the strategy is based on a typical microarray experiment with samples belonging to two classes, such as *control* vs. *affected* or tumors resistant to drugs vs. sensitive to drugs. According to their differential expression values, genes are ordered in a ranking list called L. Then, the main challenging task consists in giving biological sense to that list.

In order to do this GSEA requires a set of genes of interest, called S, to be defined prior to the execution of the algorithm. These genes are generally those belonging to a particular pathway. The aim of the GSEA is to determine whether the members of S are randomly distributed in L or are mostly at the beginning or end of the list, so as to conclude whether the pathway is differentially expressed. The algorithm follows three main steps:

Step 1: Enrichment Score (ES) Calculation

This score represents the degree to which the set S is overrepresented at the ends (top or bottom) of the list L. It is calculated by going over L, incrementing a statistical sum when a gene in S is found, and decrementing when a gene that is not in S is found. This measuring procedure corresponds to Kolmogorov-Smirnov weighted statistics [14].

Step 2: Estimation of ES's Significance Level

This factor is estimated using a procedure of permutation tests based on phenotype, which preserves the structure of complex correlations present in gene expression data. Specifically, phenotype labels are permuted and ES is recomputed for permuted data, that is, a null distribution for the ES is generated. Afterward, the p-value is calculated using this null distribution. Then, what really matters is the labels' permutation rather than the genes' permutation, so that the gene to gene correlations can be maintained resulting in a reasonable biological evaluation.

Step 3: Adjustment for Multiple Hypotheses

When a database of gene sets is evaluated, the significance level is adjusted. First, ES values for each gene set are normalized considering the size of the set, thus obtaining a normalized score called NES. Then, the proportion of false positives is controlled by means of a calculation of the False Discovery Rate [15, 16] that corresponds to each NES.

GSEA is used a lot and it is widely criticized. It belongs to the group of Functional Class Score methods [17], since it basically considers a group of genes or a pathway as a class and applies a score. The main limitation to this method is that the groups of genes or the pathways are measured independently from one another, i.e. it does not use the relations between routes or gene sets for their biological interpretation [18]. Moreover, each gene in the group is measured with the same weight, without considering that some of them might be more important in the routes. The following two methods propose an improvement to this drawback, by taking into account the relations between sets of genes, thus improving the significance of the results.

SEPEA

SEPEA, for "Structurally Enhanced Pathway Enrichment Analysis", suggests considering the significance of a given

gene in the pathway by means of the implementation of a Heavy Ends Rule (HER) [19]. This rule gives more weight to those genes that are at the beginning or end of the biological route. Moreover, it implements a Distance Rule (DR) that also gives more importance to the genes that are closely connected and that follow a flow in the pathway. In this way, the manner of giving the score to the pathway is modified with respect to that of GSEA, now adding two components according to the new rules:

- A first new component, related to the HER rule, takes a high value when a combination of the genes that are more important in the route are differentially represented in the experiment.
- The second component is based on the DR rule and takes a high value when the differentially expressed genes are closely connected or are located very close to each other in the pathway.

These two scores are then normalized and added. Normalization is carried out by calculating the median and standard deviation. Once the final score is obtained for each pathway, a p-value is computed by randomizing the data.

The study and experimentation revealed by the article demonstrates that SEPEA is very competitive with the current approaches that aim at detecting differentially expressed pathways, and that it also yields biologically relevant results. However, the analysis is performed at the level of one pathway, and it still does not pursue the objective of finding relations between several pathways in order to build a network.

SPIA

Signaling Pathway Impact Analysis [18] combines the evidence obtained from the classical enrichment analysis with a novel type of evidence, that measures the actual perturbation on a given pathway under a given condition. It proposes a new approach based on a value that is called Impact Factor (IF). IF is obtained for every pathway, incorporating other topological features. It is calculated as an addition to the next two terms:

Step 1:

A probabilistic term captures the significance of the pathway from the perspective of the set of genes that it contains.

Step 2:

A second term depends on the identification of the genes that are differentially expressed, considering the interactions described in the topology of the pathway. This term is, in essence, the sum of the perturbation factors of the genes. These factors include both the genetic information contained in the experiment and the importance of the gene in the pathway. The IF is then normalized according to the amount of differentially expressed genes.

Even though this method manages to use the information of the internal topology of the pathway successfully for a better understanding of the experimental results, one has to be reminded at this point that it still does not analyze the relations between pathways.

DEAP

This method, called Differential Expression Analysis for Pathways [20], aims at identifying pathways that are relevant in a given dataset. In order to do so, it considers the information in KEGG about the links between genes in the pathways. Then, the method represents the pathways as directed graphs, as KEGG does, and separates each pathway into simple paths or simple cycles with no repeated nodes. These paths are calculated with the absolute maximum running sum score, where the edge or relation type is signified by a representative sign summand, positive or negative. Then, the objective is achieved by finding co-regulated differential expression of their paths.

It is important to point out that the scores given to pathways are not directly comparable, as they depend on the paths that they contain. Therefore, it can be said that the score given by this method is not representative. In this regard, they try to overcome this issue by using a self-contained approach that also considers the significance of each pathway.

In comparison with the previous methods that perform the analysis at a gene level, DEAP and all other methods that use the graph structure of the pathway add a significant amount of information. This extra information is contained on the edges and is of vital importance for biologists. In most cases, these relations can make a result clearer for its study.

2.2. Inference of networks of Pathways: From Sources Different from Microarray Data

In this section, several methods that actually find associations between pathways from different sources, such as databases, annotations and the internet, are presented. These relations between pathways are often called crosstalk and they constitute a means for building pathway networks.

2.2.1. Inference Using Statistical Analysis

Function Based Analysis (FBA)

In this method [21] the proposal is to identify relations between pathways by analyzing functional links between them, based on Gene Ontology (GO) [22]. The approach is based on functions that aim at identifying some functional similarity between the routes. The information used by the authors is obtained from the database PID [23].

In GO, most genes are assigned terms or annotations that are generally based on the biological pathways of which they are a part. For this reason, annotations of a route can be inferred from its components. The proposal of FBA mainly consists of two steps:

Step 1:

During the first step, the inference of the terms that are representative for each route is carried out. This is useful because the components can have many GO terms, and some of them might not be relevant for describing the function of the pathway. The Fischer test is used in order to identify the enriched terms of a set of genes in a route.

Step 2:

Next, the similarity between routes is calculated. In principle, it can be assumed that related pathways should

share GO annotations, but this is not the only factor considered by this method. It also takes into account the content of the shared term and the number of terms in both pathways.

The main strength of this method lies in the fact that it does not only use statistical similarity, but also considers curated biological information. The main drawback, however, is that as in most methods it is impossible to find any relations between pathways that do not share any gene.

2.2.2. Inference using Data Mining

Alternatively, there are some algorithms that use data mining to find relationships between biological pathways and using this information they build a network of pathways. These algorithms aim to overcome the problem of rapid growth of publications and databases, which makes it difficult to work with updated information. The same applies to networks of biological pathways; each new publication may have found a new relationship between two or more biological pathways or as already mentioned, the limits of the pathways are not well defined, so it is really more valuable to work with the most modern information as possible.

Arizona Relation Parser

Arizona Relation Parser (ARP) mainly differs from other methods that implement data mining by using a hybrid syntactic-semantic grammar [24] for the inference of pathways. The syntactic part consists of Part of Speech (POS) labels with some other information. The semantic part is composed of specific domain words and patterns and is incorporated through a template. Thus, in this approach, the syntactic and semantic analyses are applied together, by using a larger number of labels or word classes to reflect the relevant features of the words. This implies that the rules must be defined specifically for each class, therefore it is necessary to write numerous rules in order to support all the labels. The parser was trained using 40 PubMed abstracts and then tested using 100 unseen abstracts, half for precision and half for recall.

A drawback of this approach is the need of a molecular biology expert to describe the rules for each word class before the execution of the algorithm since a great amount of biological knowledge is necessary in order to obtain significant results.

PANTex

This approach is similar to the previous one, but it only performs a syntactic analysis for the inference of pathway networks [25]. Another difference to the former method is that PANTex analyses full texts instead of abstracts, and works over all PubMed and not on a partial corpus.

As the starting point, PANTex uses KEGG to gather a list of pathways for each organism; initially, only humans and yeast are considered as valid organisms. On the base of this list, a search is performed using Entrez Utilities (NCBI Resource Coordinators, 2013) and PubMed for each pair of routes. The resulting information is stored in an intersection matrix called IRPM (for Intersection Results Pathway Matrix). Columns and rows of IRPM are pathways, and the numbers in each cell are those calculated in the previous search procedure. Also, a search of each individual route is performed, thus yielding the Pathway Results array, in order to use the values therein for normalization purposes.

The method is validated using the data reported by Alexeyenko and Sonnhammer [26]. As a conclusion from the experiments presented, it can be said that the quality of the results strongly depends on the consulted corpus. Moreover, a major problem that distorts the inference process in these kinds of strategies is the lack of standardization of the names of the pathways.

2.3. Inference of Networks of Pathways: From Microarray Data

There are some methods that not only seek topological links between biological pathways but also benefit from the information in a microarray experiment to be used when modeling those connections. The approaches referred to in this section try to tackle these problems together.

Significance Analysis of Links (SAL)

In this study, the authors [27] aimed to overcome the problems of significance analysis at the single gene level. In order to do that, they propose a method that allows the significance threshold applied to genes to be relaxed which leads to better biological knowledge. The first step consists in the selection of significant genes from a microarray experiment, using ANOVA; then they assemble a pathway network for reporting the results, so the noise of the data is reduced.

For the network, they use a single pathway as a node with an associated feature that indicates the over or under-representation according to the significance analysis gene-wise. The same classification is used for crosstalk between pathways, indicating the over or under-representation of the link by taking account of the shared genes between the respective pathways. In doing this, they show that groups of genes at the interface between different pathways can be considered as relevant, even if the pathways they belong to are not significant on their own.

The main problem with this method is that it relies on the existence of a link when the pathways have genes in common, and those genes are used to determine the significance of the link.

Method based on Protein Network

This method [28] initially builds a protein network and then, it is mapped to a pathway network. The protein network is built based on KEGG data, and it constructs the protein-protein network with microarray data by weighting the links between pathways. The proposal has 2 main steps:

Step 1:

The method first assembles the pathway network given a protein-protein network and pathway information available in KEGG. The protein-protein network is mapped to a pathway network in the following way: if a pair of pathways shares genes, then those pathways are linked.

Step 2:

Subsequently, those links are weighted by the expression and co-expression of genes in a microarray experiment, thus giving significance to the link. In this manner, the significance of a link is measured by the expression of the overlapping genes between the pathways.

The pathways and proteins used in this study have a well-known connection with Alzheimer's disease (AD) [29-31]. They use AD microarray data in six regions of the brain, as seen in the methods below. The authors also assembled a similarity matrix between pathways in order to apply clustering for further analysis. However, the method of Liu *et al*. has an important drawback as in the previous method, as the strategy is only capable of finding links between pathways if they have genes in common.

PathNet

This approach [9] uses topological information to identify associations between pathways. The main difference with the previous methods is that PathNet (Pathway-based Networks) can also find connections on pathways that do not have any genes in common. The authors also use microarray data to enrich the network with biological information. The method takes advantage of topological information (available in KEGG) and microarray data simultaneously in two principal steps:

Step 1: Pathway Network

PathNet starts by creating a set of graphs, each one representing a pathway. Each graph is a directed graph built using the information available in KEGG. The nodes of the graphs are genes and the edges are interactions in the corresponding pathway. Directed and bidirectional edges represent processes and binding events, respectively.

PathNet builds a structure named Pooled Pathway based on the pathway graphs which is reorganized for only taking into account those genes that are included in the experimental data. The information in this structure is used to create an adjacency matrix. The adjacency matrix has as many rows (and columns) as genes that are present in the pooled pathway, and stores a 1 for cell *ij* if there is a connection from gene *i* to gene *j* (for all $i \neq j$) and a 0 otherwise, as well as for all the diagonal elements. Before using the adjacency matrix on the next step, the matrix is rebuilt by deleting the rows (and columns) corresponding to genes that were not taken into account in the experimental phase (they are absent on the corresponding chipset).

Step 2: Pathway Enrichment Analysis

For the enrichment analysis step, PathNet calculates two types of information values: the "*direct evidence*", determined by the differentially expressed values when gene *i* is evaluated for two experimental conditions, and the "*indirect evidence*", which is obtained for all *i* from the values of the neighbors in the pooled pathway of gene *i*. A new piece of information called "*combined evidence*" is then generated by combining the previous ones. It is represented by a *p*-value that is used in order to determine pathway enrichment, which is done by applying the hypergeometric test to discover differentially expressed pathways and their corresponding connections.

The main contribution of this method is that it succeeds in finding connections between pathways that do not have any genes in common.

PANA

Ponzoni *et al*. have proposed another important method [10]. The approach, called PANA (from PAthway Network Analysis), uses pathway annotations in order to identify the set of genes that belongs to each pathway. Then, this method applies principal component analysis to gene expression experiment data for extracting an activity profile for each pathway. Finally, PANA infers the relationships between those profiles by using a machine learning method based on a rule-associations inference strategy. The ultimate result is a pathway network describing the functional transcriptional connections within it. To reach this result, PANA goes through two main steps:

Step 1: Pathway Compression

PANA builds a sub-matrix for each pathway by using the information available on a pathway database (in this case KEGG) and gene expression data. The method uses the database to find out which genes belong to each pathway and the expression data to obtain the expressed values for the corresponding annotated genes. When N genes are measured, only those associated with the current pathway are considered. Once this information has been collected, PANA uses bootstrapping to obtain pathway signature submatrices and reduces the profiles' dimensionality of the submatrices by applying the Principal Component Analysis (PCA) method, thus keeping the most relevant profiles (genes) that best outline the changes in the gene expression experimental data. Throughout a number of repetitions of different samples of the same data, the PCA method selects different sets of Principal Components (PCs) for each pathway. From the sum of these sets, a subset of the most frequently selected PCs is used to form a pathway signature matrix for each pathway. Finally, before going to the next step, PANA creates a Pathway Level Matrix (PLM) by concatenating the PCA score component of all the pathway signature matrices with at least one signature.

Step 2: Inference of Association Rules

Once the PLM has been built, PANA transforms its values into two possible states and obtains a discretization of the PLM matrix, named **δ**PLM. This discretization is calculated at a row level (i.e. on the pathway signature dimension). When discretizing row j, PANA considers all the values of the pathway signature k (k#j) as feasible threshold discretization values to find which one minimizes the partition entropy metric. This metric is used to evaluate how good a threshold value is when comparing the discretized k row, PLMk that results from using a candidate value, and the discretized j row, PLMj, calculated by using its mean value. Finally, PANA uses the **δ**PLM matrix to infer pathway-pathway association rules. During this step, the algorithm establishes the pathway-pathway covariation for each pair of pathways, together with an accuracy level. The pathway-pathway covariation allows the identification

of pathways that are related to each other. This task is carried out by means of a combinatorial optimization learning method, named GRNCOP [32], while the accuracy rule is measured by using sensitivity and specificity metrics. The final selected rules are those with an accuracy level higher than a predefined threshold. Then a network model is assembled using these rules.

The main strength of this method is that it finds relations between pathways even when they do not have any genes in common. Another important feature is that the method is not attached to any kind of topological information. It simply considers the activation of a pathway as a coordinated and relevant change in the expression of some of the genes and in this way they define the pathway profile.

PET

PANA and PathNet are two methodologies that detect connections between pathways even when they do not have any genes in common. PET (from crosstalk Pathway inference by using gene Expression data biclustering and Topological information) [33] is a method that also detects connections under the same circumstances, but as an additional feature, it also includes information about the synchronization between differentially expressed genes as part of the inference process. In the paper, the authors propose the use of the biclustering of gene expression data, in combination with topological analysis, in order to extract synchronized pathway associations. The method is built combining the topological information available in KEGG as a matrix with each topological link between pathways, and an enrichment analysis at the gene level given by microarray data; this analysis is used to determine whether a pathway is being active or not in the experimental data. Whenever a pathway is not active, the links associated with it are not considered. The information from both sources, active genes and pathways from the enrichment analysis, is combined using topological links as a starting point and is poured into a biclustering algorithm called BiHEA [34] in a final phase. From this algorithm, a score is computed that indicates how strong a relation between pathways is. The final result is a network at the pathway level with weighted edges.

When this algorithm is compared with the previous ones (PANA and PathNet), improvements in the amounts of relations are found. Moreover, PET provides more information about the links being found by forming biclusters that might explain synchronization between the genes that are involved within.

3. KEY REMARKS

In this article different methodologies currently used for inferring biological networks of pathways are presented, which either discover some knowledge from the overwhelming amount of biological data or attempt to find interconnections between these routes. The first three methods that were presented do not find any networks of pathways directly, but they might well be used for that aim as they detect routes that are differentially expressed simultaneously. A common disadvantage of those methods is that they treat all the genes in the pathway equally, independently of the importance of the genes in the process.

On the other hand, most of the methods that are used to find some kind of relation between pathways share one characteristic: they only consider the genes in common between pathways in order to infer the relations. This feature is common to most of the methods here presented, except for PathNet and PANA. Likewise, methods were classified according to whether or not they start from information from microarray experiments and some of them also combine this data with topological material. So, considering all these factors, we present a chart of the main characteristics of the methods presented in sections 2.2 and 2.3 in Fig. **1**.

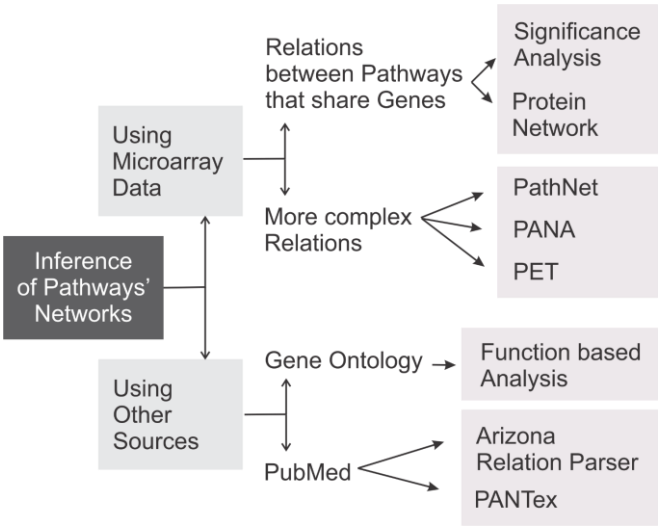

Fig. (1). Categorization of the inference methods presented that are used to build pathway networks.

Furthermore, in Table **1**, we list all the methods and associate them with their sources of information. The first column indicates the name of the method as cited above; the following columns are the different materials used in each method. If a method uses more than one source, all of them are marked in the corresponding column of the table.

4. DISCUSSION AND CONCLUSIONS

During the last decade, pathway analysis has become a key strategy for understanding the biological meaning of a gene set in high throughput experiments. For the majority of complex cellular phenomena, explanation at the gene level is hard to achieve. In order to overcome this difficulty, many methods have been proposed for identifying pathways that are enriched or differentially expressed under some specific condition.

Yet, some of the existing methods consider pathways as isolated entities in a cell, without contemplating crosstalk between them. In this context, a crosstalk refers to a situation in which one or more components of a pathway affect another, thus resulting in coordination between different biological processes. So, the methods explained in this paper overcome this limitation and explore the concept of associations between pathways using microarray experimental data and/or topological or curated information. The approaches also differ in the manner in which they deduce a crosstalk between pathways: in most cases, they only consider a link between two pathways when they have genes in common. However, there are some methods that

Table 1. Sources of information used by each method.

METHOD	Gene Set	GO	KEGG	PubMed	Microarray data
GSEA	X				
SEPEA	X		X		
SPIA	X		X		
DEAP			X		X
FBA		X			
Arizona Parser				X	
PANTex			X	X	
SAL			X		X
PP Network			X		X
PathNet			X		X
PANA			X		X
PET			X		X

search the crosstalk between pathways in a more complex way. We decided to arrange the most important methods for pathway networks inference according to the data they use as well as the findings they provide.

As has been said before, there are approaches that only find the enriched pathways, meaning that they *map* the information given by the expressed genes to a pathway level, in order to better understand the results by knowing which biological processes are activated under a specific condition. These methods are different in the way in which they consider whether the pathway is enriched or not. For instance GSEA determines statistically if the majority of the genes in the given pathway are differentially expressed or not; while SEPEA and SPIA give more importance to some genes within the pathway: SEPEA does it with the use of HER, in order to determine if a pathway is activated by the genes that drive its behavior, while SPIA does it by the use of an IF calculated for the pathway, the importance of each gene in each pathway depends on its role and its connections. On the other hand, DEAP uses directed graphs to separate each pathway into simple cycles with no repeated nodes, and then calculates the absolute maximum running sum score.

Regarding the motivation of this paper, the most interesting strategies are those that find crosstalk or links between pathways. Among these approaches, there are some that do not use microarray data, which means that they do not find any enriched pathways or links, they only assemble a pathway network. The FBA is an example of this case, the authors only use information from GO to assemble annotations at a pathway level according to the genes within, and if the pathways share a statistically important annotation, then a link between them is made. On the other hand, there are methods that use data mining in order to find links or crosstalks between pathways; these approaches try to overcome the huge amount of information available and want to take advantage of that. On this matter, we show two different methods in this paper: the Arizona Parser and PANTex. The first one provides a hybrid syntactic-semantic grammar that allows the creation of gene networks through a training corpus in order to redefine pathways or find connections between them. PANTex, on the other hand, is proposed for finding connections between pathways according to the appearances and co-appearances in a given corpus.

Lastly, there are methods that find crosstalk using microarray data, which means that they find relevant cross-talks or enriched links and pathways under a given condition. Among these methods are those that only find links between pathways if they have genes in common, such as in the SAL approach where the importance of the link is given by the differential significance of the genes shared by the pathways. Other approaches, such as PANA, PathNet and PET, do not rely on the premise that pathways have to share genes in order to be coordinated or to have a cross-talk; these methods find links between pathways that do not have any genes in common. PathNet does it through topology information from KEGG, assuming that a pair of pathways has a link if they share genes or the genes in one of them have a topological connection with the genes on the second pathway. On the other hand, PANA uses a machine learning method to find connections between the profiles of the pathways. For its part, PET finds more relations and provides extra information by yielding biclusters, which explain some kind of synchronization between the genes. In all cases, PANA, PET and PathNet achieve a pathway network that is enriched with microarray data.

At this point, an issue that should be carefully taken into account is related to the use of KEGG and similarly defined pathways in some of the cited methods. It is important to consider and have in mind that these methods can suffer from some drawbacks as their knowledge coverage on pathways is far from being complete since they “can see” only what is available and therefore they may miss some remaining information (usually the majority of differentially expressed genes).

In summary, the identification of different kinds of associations between pathways is playing a central role in systems biology, revealing information which is undetectable at a gene level. Therefore, a comprehensible understanding of the benefits and limitations of these approaches could be the key to the development of new computational strategies for genome-wide analysis.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas [Grant number 112-2012-0100471] and Secretaría de Ciencia y Tecnología (UNS) [Grant numbers 24/N032, 24/ZN26, 24/N042].

REFERENCES

[1] Tham WS, Wong SK, Mohamad MS, *et al.* A Review of Gene Selection Tools in Classifying Cancer Microarray Data. Curr Bioinform 2015; 10.

[2] Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102: 15545-50.

[3] Werner T. Bioinformatics applications for pathway analysis of microarray data. Curr Opin Biotechnol 2008; 19: 50-4.

[4] Gomez-Cabrero D, Abugessaisa I, Gisel A, *et al*. Data integration in the era of omics: current and future challenges. BMC Syst Biol 2014; 8(Suppl 2): I1. doi: 10.1186/1752-0509-8-S2-I1.

[5] Croft D, Mundo AF, Haw R, *et al.* The Reactome pathway knowledgebase. Nucleic Acids Res 2014; 42: D472-7.

[6] Jacobo-Velázquez DA, González-Agüero M, Cisneros-Zevallos L. Cross-talk between signaling pathways: the link between plant secondary metabolite production and wounding stress response. Sci Rep 2015; 5: 8608. doi: 10.1038/srep08608.

[7] Grapov D, Wanichthanarak K, Fiehn O. MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. Bioinformatics 2015; 194. 31(16): 2757-60. doi: 10.1093/bioinformatics/btv194.

[8] Chong CK, Mohamad MS, Deris S, Shamsir MS, Choon YW, Chai LE. A Review on Modelling methods, Pathway Simulation Software and Recent Development on Differential Evolution Algorithms for Metabolic Pathways in Systems Biology. Curr Bioinform 2014; 9(5): 509-21.

[9] Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. Source Code Biol Med 2012; 7: 1.

[10] Ponzoni I, Nueda MJ, Tarazona S, *et al*. Pathway network inference from gene expression data. BMC Syst Biol 2014; 8: 1.

[11] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000; 28: 27-30.

[12] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014; 42: D199-D205.

[13] Irizarry RA, Wang C, Zhou Y, Speed TP. Gene Set Enrichment Analysis Made Simple. Stat Methods Med Res 2009; 18(6): 565-75.

[14] Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 1967; 62(318): 399-402.

[15] Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001; 125(1): 279-84.

[16] Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19(3): 368-75.

[17] Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004; 20(1): 93-9.

[18] Draghici S, Khatri P, Tarca AL, *et al.* A systems biology approach for pathway level analysis. Genome Res 2007; 17: 1537-45.

[19] Thomas R, Gohlke JM, Stopper GF, Parham FM, Portier CJ. Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure. Genome Biol 2009; 10: 1-15.

[20] Haynes WA, Higdon R, Stanberry L, Collins D, Kollker E. Differential expression analysis for pathways. PLoS Comput Biol 2013; 9(3): e1002967. doi: 10.1371/journal.pcbi.1002967

[21] Hsu CL, Yang UC. Discovering pathway cross-talks based on functional relations between pathways. BMC Genomics 2012; 13: S25.

[22] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* Gene Ontology: tool for the unification of biology. Nat Genet 2000; 25(1): 25-9.

[23] Schaefer CF, Anthony K, Krupa S, *et al.* PID: the Pathway Interaction Database. Nucleic Acids Res 2009; 37: 674-9.

[24] Mcdonald DM, Chen H, Su H, Marshall BB. Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. Bioinformatics 2004; 20(18): 3370-8.

[25] Dussaut JS, Cravero F, Ponzoni I, Maguitman AG, Cecchini RL. PaNTex: A novel methodology to assemble Pathway Networks using Text Mining. In: VCAB2C. Bariloche, Argentina; 2014. p. 38-41.

[26] Alexeyenko A, Sonnhammer E. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res 2009; 19: 1107-16.

[27] Francesconi M, Remondini D, Neretti N, *et al.* Reconstructing networks of pathways via significance analysis of their intersections. BMC Bioinformatics 2008 9(Suppl 4): S9. doi: 10.1186/1471-2105-9-S4-S9.

[28] Liu ZP, Wang Y, Zhang XS, Chen L. Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains. BMC Syst Biol 2010; 4 (Suppl 2): S11. doi: 10.1186/1752-0509-4-S2-S11.

[29] Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 2002; 297(5580): 353-6.

[30] Goedert M, Spillantini MG. A century of Alzheimer's disease. Science 2006; 314(5800): 777-81.

[31] Wolfe MS. Presenilin and gamma-secretase: structure meets function. J Neurochem 2001; 76(6): 1615-20.

[32] Gallo CA, Carballido JA, Ponzoni I. Discovering Time-lagged rules from microarray data using gene profile classifiers. BMC Bioinformatics 2011; 12: 1-21.

[33] Dussaut JS, Gallo CA, Cecchini RL, Carballido JA, Ponzoni I. Crosstalk Pathway Inference using Topological Information and Biclustering of Gene Expression Data. Biosystems 2016; 150:1-12. doi: 10.1016/j.biosystems.2016.08.002.

[34] Gallo CA, Carballido JA, Ponzoni I. BiHEA: A Hybrid Evolutionary Approach for Microarray Biclustering. Lect Notes Comput Sci 2009; 5676: 36-47.