Computational and Structural Biotechnology Journal 21 (2023) 378–387

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

PathIN: an integrated tool for the visualization of pathway interaction networks☆

George Minadakis [a,*], Kyproula Christodoulou [b], George Tsouloupas [c], George M. Spyrou [a]

[a] Bioinformatics Department, The Cyprus Institute of Neurology & Genetics, 6 Iroon Avenue, 2371 Ayios Dometios, Nicosia, Cyprus | PO Box 23462, 1683, Nicosia, Cyprus
[b] Neurogenetics Department, The Cyprus Institute of Neurology & Genetics, 6 Iroon Avenue, 2371 Ayios Dometios, Nicosia, Cyprus | PO Box 23462, 1683, Nicosia, Cyprus
[c] HPC Facility, The Cyprus Institute, 20 Konstantinou Kavafi Street, Aglantzia, 2121, Nicosia, Cyprus

ARTICLE INFO

Article history:
Received 16 September 2022
Received in revised form 16 December 2022
Accepted 16 December 2022
Available online 18 December 2022

Keywords:
Pathway Networks
Database repositories
Pathway analysis
Graph theory

ABSTRACT

PathIN is a web-service that provides an easy and flexible way for rapidly creating pathway-based networks at several functional biological levels: genes, compounds and reactions. The tool is supported by a database repository of reference pathway networks across a large set of species, developed through the freely available information included in the KEGG, Reactome and Wiki Pathways database repositories. PathIN provides networks by means of five diverse methodologies: (a) direct connections between pathways of interest, (b) direct connections as well as the first neighbours of the given pathways, (c) direct connections, the first neighbours and the connections in between them, and (d) two additional methodologies for creating complementary pathway-to-pathway networks that involve additional (missing) pathways that interfere in-between pathways of interest. PathIN is expected to be used as a simple yet informative reference tool for understanding networks of molecular mechanisms related to specific diseases.

1. Introduction

Pathway-based analysis has become a fundamental and rapidly accumulated field of research aiming to increase the capacity to explore large-scale omics data through the understanding of the molecular mechanisms, involved in complex diseases [1,2]. The plethora of available pathway databases has raised the need for developing tools and repositories, to integrate this type of knowledge into a more holistic representation that involves network-based approaches [2,3]. Online tools for the visualization and analysis of biochemical pathways are under development for over two decades. At a gene and protein level analysis, eminent works like the STRING repository [4] are in a position to provide very informative networks that depict the functional associations between proteins, thus facilitating the analysis of modularity in biological processes. Commonly used enrichment tools like EnrichR [5],GeneTrail [6,7], and gProfiler [8] can efficiently provide ranked lists of pathways and strongly relate them with the biological condition under study. However, in some cases, pathway analyses may result in large lists of pathways that are difficult to be translated and related to a specific disease without network visualisation. Moreover, strict thresholding methodologies may lead to small and incomplete lists of significant pathways associated with the biological condition under study. At the same time, no straightforward post-enrichment tool exists that visualizes at network level pathways related to a particular gene, compound, reaction, and pathway of interest, and further proposes additional pathways that may be missed from the standard enrichment analysis. Thus, it remains difficult to answer fundamental questions, such as: (a) which are the most closely neighbouring pathways that may be related to a single pathway of interest, (b) what is the connecting distance between two pathways, and which pathways are involved in-between them, (c) which pathways involve specific genes, or compounds or reactions of interest, and (d) how someone could easily perform additional methodologies to bring (or not) missing and critical pathways into their study. Expanding on this concept at a pathway level, we introduce PathIN: a generalized

☆ PathIN is available at: https://bioinformatics.cing.ac.cy/pathin, and https://pathin.cing-big.hpcf.cyi.ac.cy.

* Correspondence to: George Minadakis, Bioinformatics Department, The Cyprus Institute of Neurology & Genetics, 6 Iroon Avenue, 2371 Ayios Dometios, Nicosia, Cyprus | PO Box 23462, 1683, Nicosia, Cyprus.
E-mail addresses: georgem@cing.ac.cy (G. Minadakis), georges@cing.ac.cy (G.M. Spyrou).

https://doi.org/10.1016/j.csbj.2022.12.028

web-tool that provides an easy and flexible way for rapidly relating pathways together, based on fundamental issues related to complex networks and graph theory. PathIN holds a large database repository of reference pathway networks across a large set of species, which have been developed through the freely available information included in the KEGG [9], Reactome [10] and Wiki Pathways [11] database repositories. Through this web-service, users can easily import (or search for) pathways, genes, compounds, reactions and enzymes of interest to be translated into a network (graph) of pathways. Using network approaches, PathIN further provides five diverse methodologies for network creation that the user can select. These are: (a) pathway networks that depict all the direct connections between pathways of interest, (b) pathway networks that include the direct connections as well as the first neighbours of the given pathways, (c) pathway networks that include the direct connections, the first neighbours and the connections in between them, (d) complementary pathway networks that include additional (missing) pathways that interfere in-between pathways of interest, and (e) an extended approach for the creation of complementary pathway networks. The application further allows the calculation of network statistics based on the degree, closeness, centrality and betweenness measures derived from graph theory. PathIN is expected to be used as a reference post-experimental tool to understand the functional environment around selected molecular mechanisms of interest within the framework of network medicine.

2. Software description and methods

2.1. The pathway reference network repository

The pathway-to-pathway network draws information from a web-service that holds a large database of reference pathway networks, which have been developed through the freely available information included in the KEGG [12], Reactome [13] and Wiki Pathways [11] database repositories. Herein, the functional relation between two pathways, as provided by these databases, is considered when the first pathway involves or is involved into the second pathway, accordingly. A pathway that involves a pathway may refer to any kind of relation between two pathways, subjected to the way in which these repositories have structured their pathway relations. In effect, this type of information can form an undirected-unweighted pathway-to-pathway network. In this line of thought, we retrieved pathway information for all the organisms included in the three repositories mentioned above, parsing all the available information about the functional connections between the available pathways. Specifically, we obtained pathway information for 177 organisms (species) from KEGG, 16 from Reactome and 38 from Wiki Pathways repositories. This resulted to totally 231 undirected pathway reference networks, one for each specie, which in turn were merged in to one large network per repository. The following table shows in details these amounts per repository. Tables 1–4.

An extra-layer above the aforementioned binary connectivity between the pathways is offered using the edge weight attributes. The following table shows the supported reference network types according to the terminology of each repository.

Herein, the edge weight between two pathways was formed through the number of common biological entities, such as: genes, compounds, reactions and pathways. In the same manner, the total number of each of these biological entities can also form the node size of the pathway, accordingly. For example, if the user selects to work with genes, then the edge weights represent the number of common genes the two pathways share, while the node sizes represent the number of the genes involved in the pathway. The same approach is used in the case of compounds, reactions, metabolites and pathway sharing options. However, pathway connectivity does not necessary means commonality of elements. The relation between two pathways may also be a single membership, namely a relation without necessary sharing any specific element. In that case the edge weight is zero, and the connection remains in the network since this cannot be ignored. On the contrary, if two connected pathways share the same pathway according to their reference map, then the edge weight in between them is 1.

Table 1
Information content of the reference network repository.

Repository	Num of Species	Total Pathways	Total NET Edges
KEGG	177	24878	124707
Reactome	16	20398	21198
Wiki Pathways	38	2753	1433

Table 2
The supported reference network types and terminology.

KEGG	Reactome	Wiki Pathways
Genes	Genes	Genes
Compounds	Reactions	Metabolites
Pathways	Pathways	Pathways

2.2. General design and implementation

PathIN comes with a web interface consisting of (a) the mainframe, which holds all the appropriate tools for the users to perform their analysis, and a detailed help page written in HTML, PHP and JavaScript. The backend of PathIN has been written in an R environment, where several functionalities have been parallelised to achieve fast performance. Users can quickly evaluate, test and understand the PathIN functionalities through several available examples, which can be automatically loaded from the web interface. The following table depicts the list of libraries and modules used for the implementation of the web interface.

PathIN algorithms work with pathway IDs in how these have been defined by each pathway repository. The overall workflow of PathIN is depicted in Fig. 1. Specifically, there are three input type formats users can easily upload or create through the available interface: (a) simple lists of pathway IDs or two column tab-separated edge lists of pathway IDs, (b) lists of gene symbols, (c) lists of compounds (or reactions/metabolites) according to the terminology of each repository. Pathway IDs are directly translated to network graphs of pathways (pathway-to-pathway networks) on which users can perform additional methodologies for further network manipulation. For the rest of the input types, a specific process identifies the pathways (pathway IDs) that involve at least one of these biological entities, which are translated into pathway-to-pathway networks. The underlying utility aims to answer which pathways involve a specific biological entity of interest and how these pathways form a network which is related to a particular biological condition or disease of interest. An auto-complete utility has also been rooted in the web-interface which allows searching either for pathways (by name) or specific compounds, reactions, and metabolites of interest. Users can interactively create their own lists of interest through this utility, bypassing the limitation of knowing a priori specific IDs.

The tool can be accessed through the webpage of the Bioinformatics Department at the Cyprus Institute of Neurology and Genetics (CING) https://bioinformatics.cing.ac.cy/pathin. It is also containerised and available at the High-Performance Computer Facility (HPCF) of the Cyprus Institute (https://pathin.cing-big.hpcf.cyi.ac.cy). Expanding on the network creation concept, we further rooted into PathIN five diverse methodologies for the extraction of pathway-to-pathway interaction networks, described in detail in the following sub-sections. The calculations used in those methodologies refer to the unweighted and undirected versions of the pathway-to-pathway reference networks.

Table 3
List of libraries and modules used for the implementation of the web interface.

Module	Languages	Packages/libraries Used
Frontend Web Interface	html, php, JavaScript	bootstrap, jQuery, datatables, vis.js
Database Repository	php	MySQL
Backend interface (algorithms, functionalities, pipelines, etc)	R	curl, igraph, dplyr, openssl, jsonlite, apcluster, pracma, foreach, doParallel, RMySQL, stringr
Network manipulation & methods	R	igraph, custom R scripts

2.3. Method 1: original pathway-to-pathway network creation

The underlying methodology aims to map pathways using the reference network repository of pathways described in previous sections. Mapping is achieved by finding only the direct connections between pathways of interest, as shown in Fig. 2, where a network of three KEGG pathways is depicted. Herein, the absence of direct connections between pathways may result in networks with unconnected pathways (nodes). For example, in the network of pathways depicted in Fig. 2a, the "regulation of actin cytoskeleton" node has no direct connection with any of the remaining pathways in the network. We start with the assumption that these three pathways are strongly related with the disease under study as a result of a well-performed pathway enrichment analysis. where a specific threshold has been performed on their score. Our hypothesis here is that since these pathways are strongly related with the disease under study then there should be a minimum connectivity between them. In this line of thought, unconnected significant pathways may indicate that there is a missing link between those pathways and the rest (connected) pathways of the network. Methodologies that follow aim to improve this limitation by including additional pathways, which in turn may form more complementary and informative networks without significantly exceeding the size of the final network.

2.4. Method 2: identifying missing pathways based on shortest-paths

As opposed to the latter approach, herein, the underlying methodology identifies and adds key pathway nodes that ensure the network's minimal connectivity. A specific script finds and calculates all the shortest paths within the reference network that interconnect the pathways of interest and chooses only those nodes that belong to the shortest path-length to be included in the final network. Herein the shortest path between two distant nodes is calculated employing the breadth-first search (BFS) algorithm [14] used in igraph R package [15], suitable for unweighted networks. The BFS algorithm is able to traverse a graph, starting from a root vertex and spreading along every edge simultaneously. Formally, the BFS algorithm visits all vertices in a graph that are k edges away from the root vertex before visiting any vertex $k + 1$ edges away. This is done until no more vertices are reachable from the root vertex. When the algorithm finds more than one shortest path of the same path-length, these additional paths are also included into the final network. An example of this outcome is depicted in the complementary network of Fig. 2b, where four additional pathways (green circles) have been added to the initial 3-node network (blue-circles) shown in Fig. 2a. These new pathways are able to interconnect the underlying unconnected pathways, by keeping the shortest path length in the network. The underlying methodology allows the production of more complementary networks where all pathways are closely connected.

2.5. Method 3: identifying missing pathways based on forced pairwise connectivity

The above described missing pathway approaches draw from a recently introduced work, namely the PathwayConnector [16,17], which in turn has been further updated to support networks obtained from PathIN. Expanding on this type of pathway networks, in the latest version of PathwayConnector, an additional process has been rooted, that allows to shortly expand the size of produced complementary networks, and bring to our attention additional pathways that may be related to the specific biological condition or disease under study. Specifically, focusing on the original network of pathways depicted in Fig. 2a, the minimal complementary network is depicted in Fig. 2b by means of the above-described methodology. However, it is observed that although the network is a connected network, there is no direct information between the pathway entitled "tgf-beta signalling pathway" and the node entitled "Regulation of actin cytoskeleton". The nodes "Pathways in cancer" and "TGF-beta signaling pathway" already keep minimal connectivity, which in effect does not allow the BFS algorithm to expand the network with additional nodes. Someone would eventually like to know which other intermediate key nodes interconnect the "tgf-beta signalling pathway" with the "Regulation of actin cytoskeleton" by slightly expanding the underlying network. In this line of thought and in the prospect of covering such a scientific question, a new algorithm was further developed that examines pairwise the shortest paths between two nodes, independently of whether the network keeps or not its minimal connectivity. Specifically, for a given graph $G = (V, E)$ with n total nodes and m input nodes, we calculate a matrix that holds the pairwise combinations of m. For each pair of nodes, we perform the BFS algorithm without taking into consideration the connectivity of the rest $(m - 2)$ candidate nodes. In effect, this process brings all the possible intermediate nodes between the specific pair under study, which are then stored in a list, until all combinations are completed. This process safely brings additional nodes and edges (if any), without significantly affecting the overall size of the network. An example of this attempt is depicted in Fig. 3 where the *"proteoglycans in cancer"* pathway along

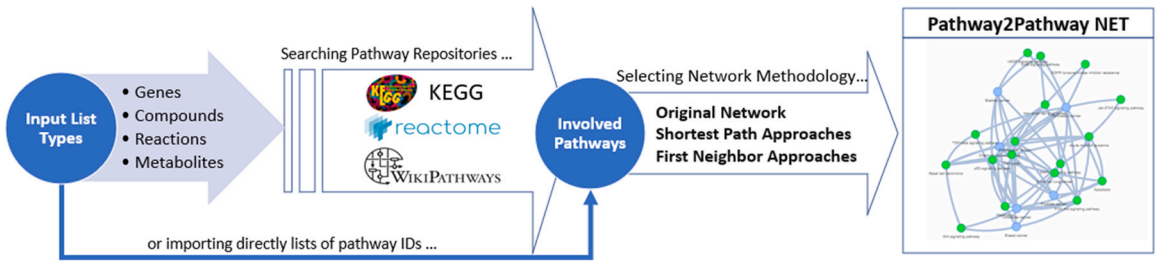

Fig. 1. Overall workflow, Figure depicts the overall workflow of PathIN web tool and its main components.

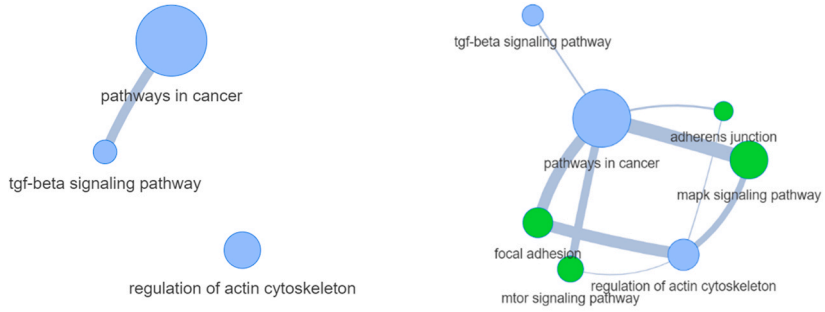

Fig. 2. Human pathway2pathway networks, (a) Original network of 3 human KEGG pathways where the edge weights characterize the number of common genes between two pathways and the functional relation between pathways. The node sizes refer to the number of genes involved in each pathway. (b) The complementary network has four additional pathways (green circles) added to the initial 3-node network (blue circles). The calculations of shortest paths in the networks refer to the unweighted and non-directed versions of the pathway-to-pathway reference networks.

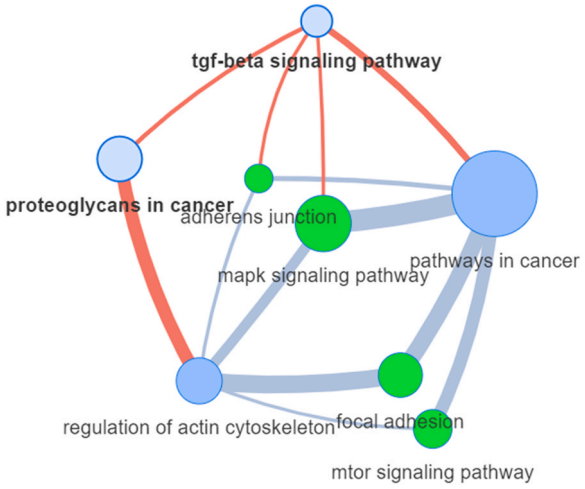

Fig. 3. Extended complementary network.

with three additional edges have been included in the complementary network, interconnecting the "tgf-beta signalling" pathway with the "Regulation of actin cytoskeleton" pathway.

Extended complementary network where 2 additional pathways (green circles) have been included to the initial 3-node network (blue circles). Herein, the edge weights characterize the number of common genes between two pathways and the functional relation between them. The node sizes refer to the number of genes involved in the specific pathway.

2.6. Method 4: identifying pathways based on first-neighbours network connectivity

We recall that the overall concept of PathIN is not only to map the pathways of interest on the reference network, but to further bring pathways in to attention that may be related to a specific disease under study. These may not be necessary pathways of high significance but pathways which are simply very close to the subnetwork of pathways under study. The above-described methodologies, have been designed to bring new pathways that interconnect other pathways of interest using shortest path approaches. However, there are also pathways which are closely related with those of interest and do not necessary act as interconnected nodes. These may be the first neighbours of a single pathway of interest which are very close to the subnetwork under study. In this line we further developed additional methodologies that slightly expand existing networks of pathways using first neighbour network approaches. Specifically, for each pathway of interest, a specific process finds all the direct connected pathways (first neighbours) and creates a pathway-to-pathway network with these nodes. The underlying method keeps only the edges between the input pathways, if any, and not the relations between first neighbours. Fig. 4a depicts an example of two pathways of interest (blue circles), which share two of their first neighbours, namely, the "metabolic pathways" and the "glycosaminoglycan biosynthesis".

2.7. Method 5: extended first neighbour approach

In the prospect to further expand small networks of pathways, we further introduce an additional methodology that slightly increases the information content of networks obtained by method 3. Herein for each pathway of interest, a specific process finds the first neighbours of each pathway under study as well as their connections in between them, resulting to an even more connected pathway-to-pathway network. In effect, this method further allows to examine the existence of intermediate and direct connections between pathways of interest, as well as the connections in between the first neighbours. Fig. 3b depicts an example of this methodology, showing that the previously mentioned network shown in Fig. 3a, is now enriched with additional edges (depicted in red colour) which in turn refer to the connections between the first neighbours.

The above-described methodologies provide innovative information content between pathways of interest, allowing users to create informative pathway networks that may be related to a specific disease or biological condition of interest. Users can easily download all the produced networks in a tab-separated text file form that holds all the information rooted in these reference networks.

2.8. Software performance: on the connectivity of spastic ataxia related pathways

Herein, the evaluation of PathIN draws from our recent work based on identifying pathways related to spastic ataxia following differential expression analysis (DEA) on 22 human microarray gene expression datasets derived from various tissues of patients with ataxia or spasticity [18]. As a part of that work, we examined whether the *"sphingolipid signalling pathway" and the "sphingolipid metabolism"* could be considered candidates for developing the disease phenotype. In this context, the top scored genes obtained from DEA analysis, were further analysed using specific tools for pathway enrichment, and pathway identification [19,20]. This analysis revealed

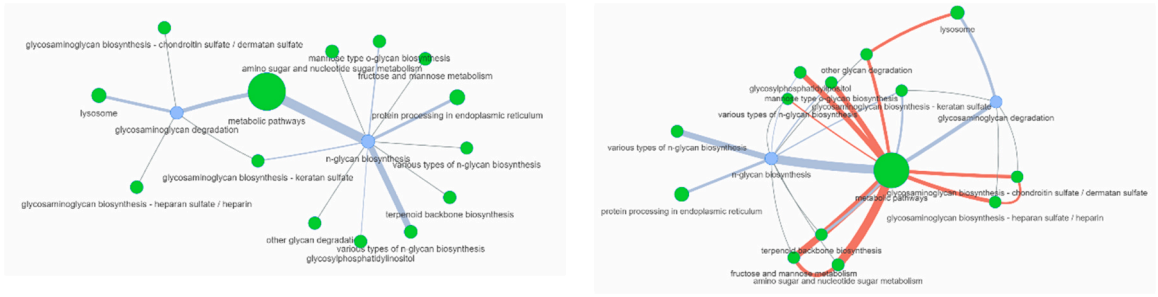

Fig. 4. Human pathway-to-pathway networks based on first neighbour approaches, (a) Network of 2 human KEGG pathways of interest and their first neighbours, obtained using the first neighbour approach. (b) Network of the same pathways of interest and their first neighbours, obtained through the extended first neighbour approach, where additional edges (depicted in red colour) refer to the connections between the first neighbours.

that the "*PI3K-Akt signalling pathway*", "*MAPK signalling pathway*", "*Calcium signalling pathway*","*cAMP signalling pathway*", "*Apoptosis*", "*AMPK signalling pathway*", and "*Insulin signalling pathway*", are common among the neuronal, peripheral blood and fibroblast "ataxia" datasets. A network of these pathways is depicted in Fig. 5, using the 1st methodology described in the previous section.

Undirected network of 2 human KEGG pathways, obtained by means of the 1st methodology described in the text. The red edges highlight the first neighbour connections around the selected node.

We further perform literature search to examine whether the methodologies used for the identification of missing pathways, can adequately lead to candidate pathways for consideration. Our first approach attempts to create an even larger network that brings additional missing pathways on one hand and keeps the minimum connectivity of the network on the other. Specifically, using the above network of pathways, we performed the 2nd and 3rd methodologies, which are based on the shortest path approach, generating a network depicted in Fig. 6a. It is observed that 8 additional pathways (depicted in green colour) have been included in the underlying network. Indicatively, the "*Phospholipase D signalling pathway*", which is one of the first neighbours of the "*sphingolipid metabolism*" has already been mentioned in studies related to ataxia [21–23], as well as pathways related to fatty acids [24,25]. Analysis using the 3rd methodology resulted in a larger network (depicted in Fig. 6b), where additional intermediate pathways bridged the connection between the "*sphingolipid signalling pathway*" and the "*sphingolipid metabolism*". Furthermore, additional pathways have also been included that interconnect the "*PI3K-Akt signalling pathway*", and the "*Calcium signalling pathway*", without necessarily suggesting a straightforward relationship with the "*sphingolipid metabolism*" and the "*sphingolipid signalling pathway*". On the contrary, the latter observation suggests that some of these pathways could potentially be used to bridge these two candidate pathways under study.

A second approach that has been employed, starts from an initial network that includes the "sphingolipid signalling" and the "sphingolipid metabolism" pathways. Specifically, we performed the 5th methodology in the prospect to expand the underlying network in terms of their first neighbours and their in-between connections. The outcome of this process is depicted in Fig. 7a, where the obtained network has been enriched with additional edges in-between these first neighbours. However, none of these first neighbours interconnect the two underlying pathways of interest, except the direct connection between them. Another observation here is that the "PI3K-Akt signalling pathway" and the "MAPK signalling pathway" mentioned in [18], have also appeared in these networks, as first neighbours of the "sphingolipid signalling" pathway. Expanding on the latter approach, the aim here is to bridge the initial 2 pathways of interest by creating an extended complementary network on one

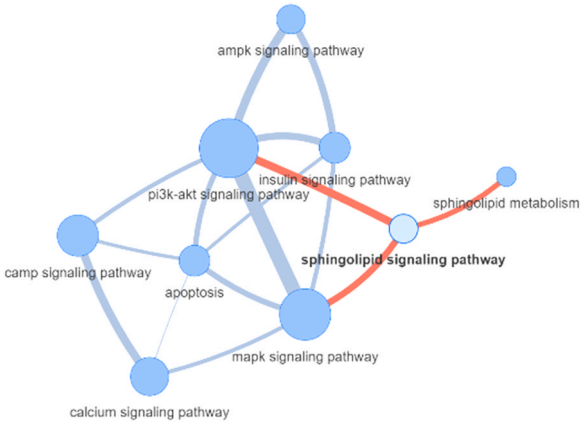

Fig. 5. Human pathway-to-pathway network.

hand and further keeping a network with minimal connectivity on the other. For this approach we used as input the obtained pathways that are depicted in Fig. 7a, on which we performed the 2nd methodology. Fig. 7b depicts the obtained network, showing that 7 additional pathways (depicted with green colour) have been further included in the network, also mentioned in studies related to ataxia and other neurodegenerative disorders. Indicatively these include the "Carbon Metabolism in Cancer" [26], "Gluconeogenesis" [27], "Glycosaminoglycans" [28], "mTOR signalling" [29,30].

Herein it should be stressed that the above analysis by no means suggests any objective truth on the underlying mechanisms related to ataxia, but just a simple biological scenario that aims to show the performance of PathIN as a tool for the investigation of pathway interactions.

2.9. Biological importance and statistical significance of implicated pathways

The importance of biology and the implicated pathways that derive from PathIN methodologies, has been initially examined in [16] through a comparative statistical analysis between enrichment analysis tool, random pathways and pathways deriving from the missing pathway methodology described in this work. The missing pathway methodologies were further performed with noteworthy results to pathways related to Alzheimer's Disease (AD) [31], to Huntington's disease (HD) and Spastic Ataxia (SA) [18,32], as well as to identify pathways related to the inhibition of Breast Cancer cell invasion [33]. However, the concept of PathIN is not to serve as a traditional pathway enrichment tool but as a post analysis tool that deals with the results obtained from any kind of pathway-based

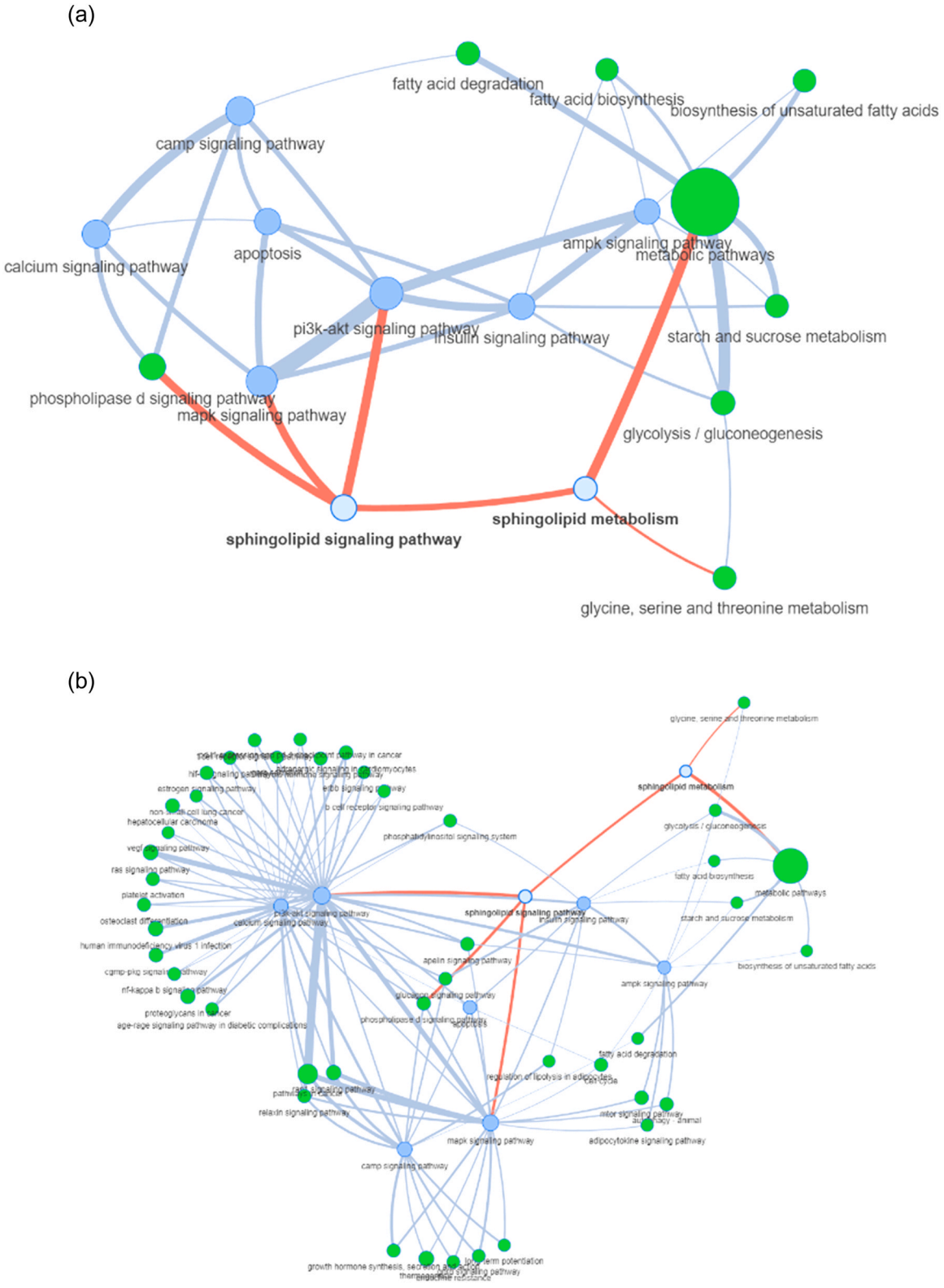

Fig. 6. Human pathway-to-pathway network (a) Network of human KEGG pathways, obtained using the 2nd method described in the text (b) Network of human KEGG pathways, obtained through the 3rd method described in the text.

analysis that brings candidate pathways related to a specific biological condition. PathIN assumes that the significance/importance of the input pathways has already been evaluated by the user and subjected to the user's biological question. PathIN doesn't deal with the significance of each pathway but it provides a network-based information about its connectivity with other pathways, by examining and suggesting a plausible subnetwork where these pathways of interest may belong to. The proposed methodologies aim to bring pathways that are not necessary top-scored on one hand, but very close to those of interest on the other. Insignificant pathways of interest are also in the scope of PathIN for those who want to address generalized biological questions. This is a crucial

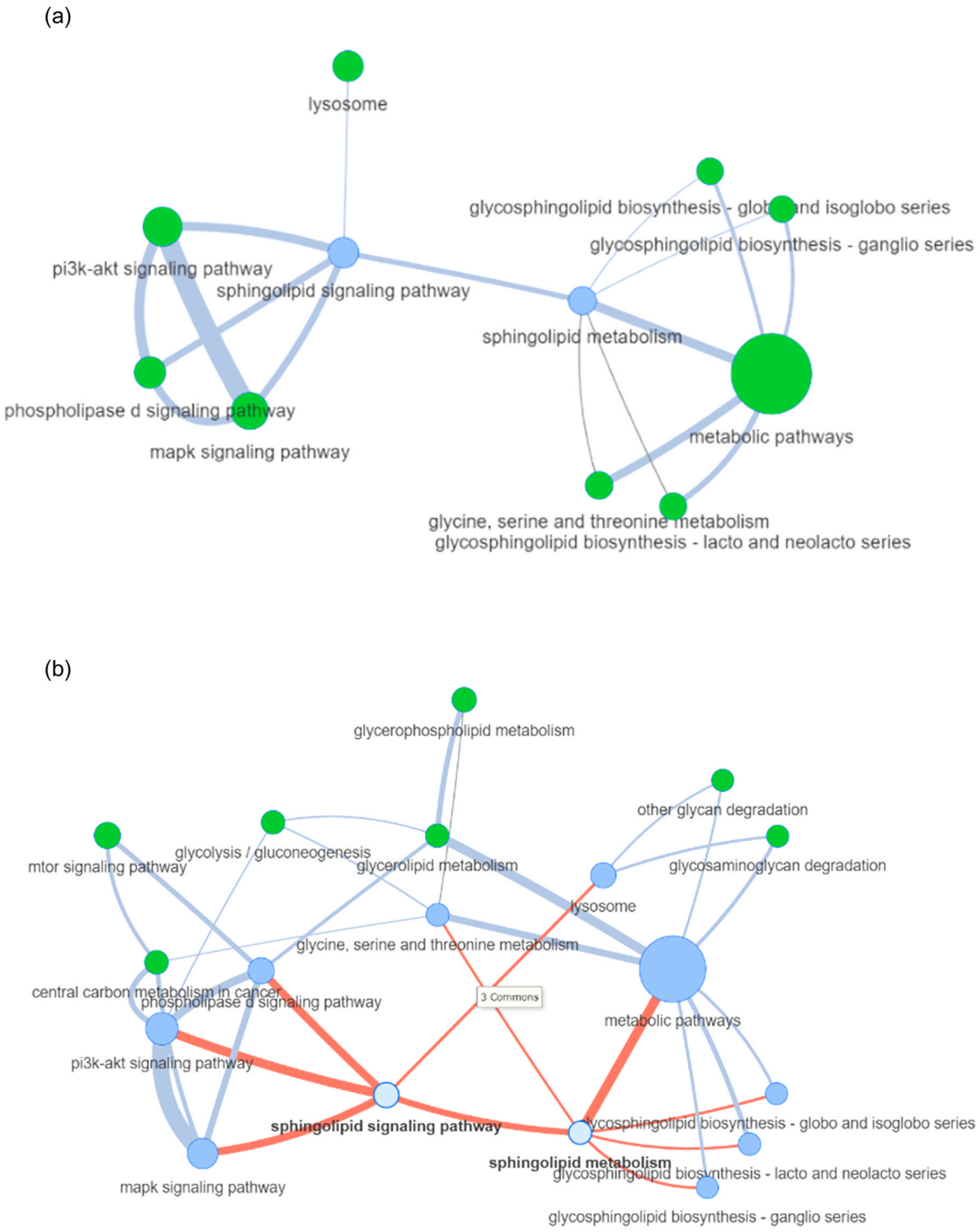

Fig. 7. Human pathway-to-pathway networks based on first neighbour approaches (a) 2 human KEGG pathways (blue circles) along with their first neighbours (green circles) (b) the same network (blue circles) expanded using the 2nd methodology described in the text.

difference in between PathIN and other tools. PathIN does not use any predictive algorithms for candidate pathways, but simply provides a well-designed/evaluated pathway connectivity which is subjected on the biological, structural and functional parameters of the repositories that draws from. However, in the following section we employ a statistical methodology in order to show that pathways which are closely connected to a significant cluster of well-grounded pathways, may also be related or relevant to the condition under study.

2.10. Statistical significance of implicated pathways using 2case studies

In practice there is not an optimal way to find a solid "ground truth", namely a set of either ranked genes or ranked pathways that are related to a disease and we understand that this is a problem when validating such kind of methods. However, in order to bypass this limitation and to provide an accepted notion of statistical significance, we further employ an additional methodology used in [16], that shows the significance of the implicated pathways that

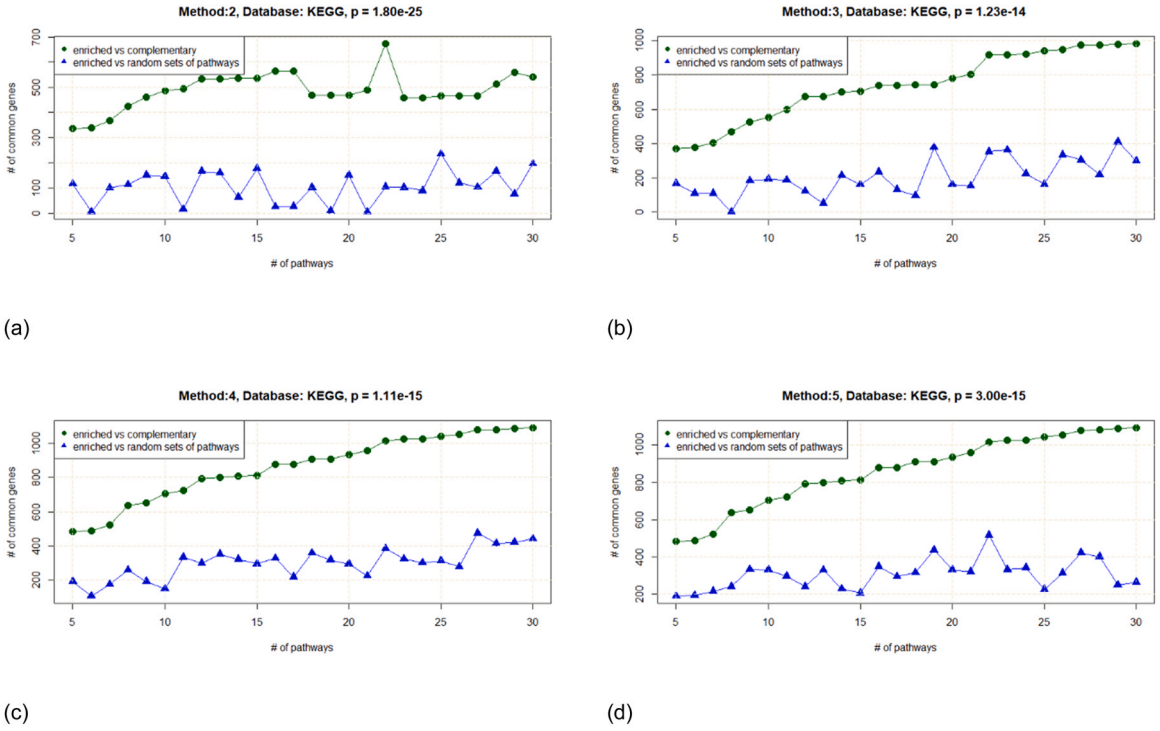

Fig. 8. Statistical analysis of implicated pathways related to sepsis Graphs depicting the significance of complementary pathways compared to randomly selected ones. Distributions show the number of common genes for different number of enriched pathways between complementary (green bullets) and random pathways (blue triangles).

derive from the introduced methodologies. Assuming that we have a list of *N* top-scored candidate genes related to a specific disease. Then, using the EnrichR [5] package, we further perform enrichment analysis to end up with a sorted list of *K* top-scored pathways. Starting from a seed of the first five top-scored pathways in the *K* list and adding one pathway at a time from the same list, we are writing down the new pathways that appear by performing the introduced methodologies. In the sequel, we are counting the number of common genes between the new pathways and the seed pathways and we compare this number with the number of common genes found between the same number of randomly selected pathways and the seed pathways. These two distributions are compared with a z-test algorithm, in order to estimate their statistical significance difference. To examine latter approach, we firstly focus on a recent study on mRNA expression data from blood samples taken from patients with sepsis and healthy individuals [34]. The authors provided a list of genes showing that AGTRAP, IRAK3, ADM, ALOX5, MMP9, S100A8 and ENTPD1 have potential diagnostic value in sepsis. Herein, we used this list of genes to perform pathway enrichment analysis by means of the EnrichR package. The outcome of this process was 37 KEGG pathways sorted by means of the *Combined Score* provided by the EnrichR tool. Starting with the first 5 top-scored pathways, we performed the above-described comparative analysis in order to examine whether the methodologies 2,3,4 and 5 described in this work can adequately bring significant and relevant pathways with those derived by an enrichment analysis tool. Fig. 8 depicts the outcome of this attempt. It is observed that the pathways deriving from those methodologies (see green bullets) have increased commonality at gene level comparing to those that derive from random selection (blue triangles). The low p-value obtained from the z-test algorithm further indicate that the two distributions have significant statistical difference. In effect this observation further shows the irrelevance of the random implicated pathways with the network of pathways under study.

In order to further examine the reproducibility of the method, in the following we focus on a different case study related to the colon cancer related genes identified in [35]. Herein the authors mention 24 candidate genes related to hypermutated and non-hypermutated cancer, obtained from a genome-scale analysis of 276 samples, analysing exome sequence, DNA copy number, promoter methylation and messenger RNA and microRNA expression data. Enrichment analysis revealed 110 KEGG pathways sorted by means of the *Combined Score*. Fig. 9 depicts the results of the statistical analysis across the four methodologies, showing the significant difference in commonality of genes between the two distributions.

Graphs depicting the significance of complementary pathways compared to randomly selected ones. Distributions show the number of common genes for different number of enriched pathways between complementary (green bullets) and random pathways (blue triangles).

Although gene-commonality does not necessary means significance or relevance for a candidate pathway, it still remains as one of the factors used by enrichment tools to score pathways. The above results indicate that when having a cluster of pathways strongly related to a disease/condition under study, neighbouring pathways to that cluster may also be relevant at gene-commonality level, despite the fact that are statistically insignificant. This is consistent with the missing pathway concept employed in this work, towards bringing closely connected complementary pathways.

2.11. Comparison with other tools

Eminent tools that involve pathway network visualisation such as ClueGO [36], PathExNET[37], PathME [38], PANEV [39], ComPath [40], and many other [41,42], can successfully create pathway-to-pathway networks. However, although these applications are based on the connectivity information provided by the available pathway repositories, the commonality numeric information draws only from the genes involved in the pathways. On the contrary, the PathIN software further allows the construction of pathway networks that draw from commonalities at compound, gene, reaction and metabolite level information, while simultaneously provides several

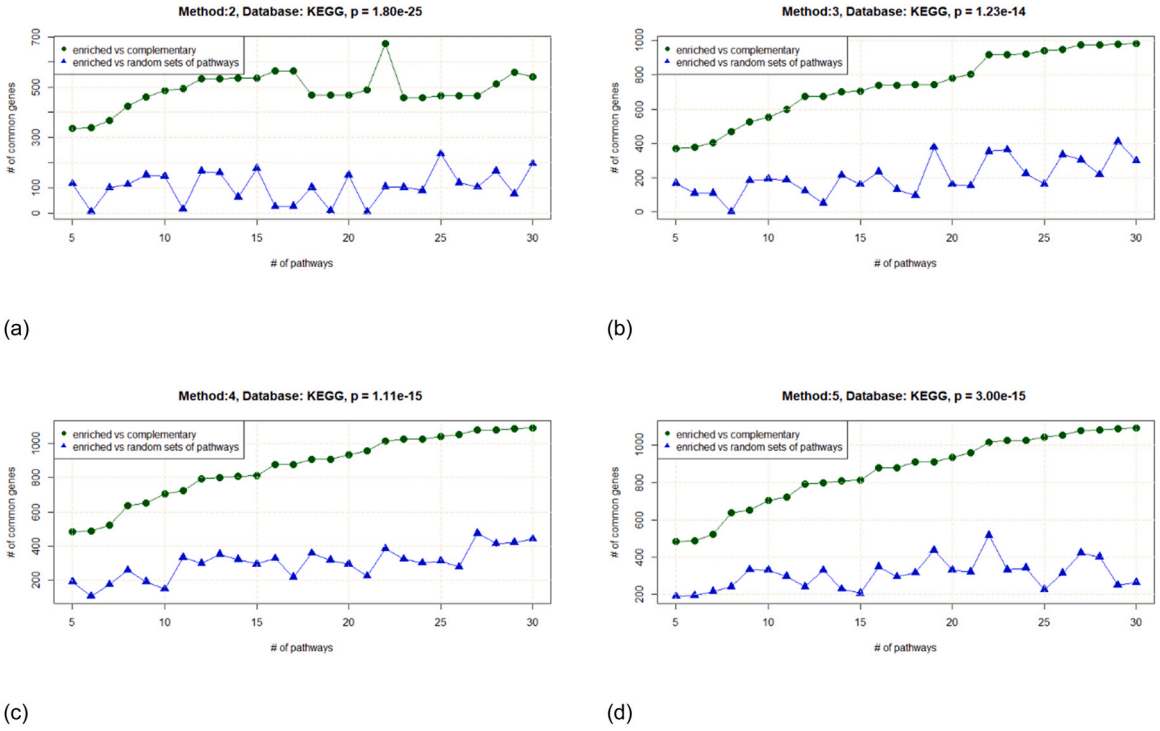

Fig. 9. Statistical analysis of implicated pathways related to colon and rectal cancer.

Table 4
A list of existing tools compared with PathIN functionalities. Herein, the aces mean that the underlying functionality is supported by the specific tool. The last column refers to an indicative score calculated by the sum of the aces.

Tool	Web	Score-Based	Multiple Species	Supported input entities for edge commonality			Supported Repositories			Overall Score
				Genes	Compounds/ Metabolites	Reactions	KEGG	Reactome	Wiki Pathways	
PathIN	1	0	1	1	1	1	1	1	1	8
PathExNET	1	1	1	1	0	0	1	1	1	7
Pathway Connector	1	1	1	1	0	0	1	1	0	6
PathME	1	0	1	1	0	0	1	1	1	6
PANEV	0	0	1	1	0	0	1	0	0	3
ClueGO	0	1	1	1	0	0	1	1	1	6

methodologies for manipulating these networks depending on the user's specific scientific question. The following table provides an indicative list of existing tools compared with PathIN functionalities.

3. Discussion

Casting biological pathways as networks has become a promising and valuable Systems Bioinformatics approach that aims to enhance post-experimental omics analyses and to provide further insights on the functional inter-relation of pathways [16,17,37]. The powerful concept of the graph theory and the plurality of pathway-based information included in eminent database repositories such as the KEGG [9], the Reactome [10] and the Wiki Pathways [11], has put significant contribution to the understanding and development of novel software that allows the integration of pathway data at network level [43,44]. The creation of pathway networks through the PathIN web-framework can act as a reference map of implicated pathways, which is a promising approach towards enhancing our understanding of the biological mechanisms that may be related to a specific biological condition under study. Novel genes, compounds and reactions derived from additional omics sources or even smaller scale laboratory experiments can potentially be projected on this map. The PathIN outcome, has been further evaluated and verified in this work, through a comparative statistical analysis between outcomes of enrichment analysis tool, and implicated pathways deriving from PathIN methodologies. PathIN holds an integrated environment for pathway connectivity that aims to act as a post-analysis tool, providing additional information on the candidate pathways obtained from traditional pathway analyses.

CRediT authorship contribution statement

George Minadakis and George M. Spyrou carried out the implementation and design of the proposed web-tool. Kyproula Christodoulou evaluated the tool due to her expertise in specific case study. George Tsouloupas containerised the tool to be served through the High-Performance Computer Facility (HPCF) of the Cyprus Institute.

Conflict of Interest

The authors have declared no Conflict of Interest.

References

[1] Jin L, Zuo X-Y, Su W-Y, Zhao X-L, Yuan M-Q, Han L-Z, et al. Pathway-based analysis tools for complex diseases: a review. Genom Proteom Bioinform 2014;12:210–20.

[2] Oulas A, Minadakis G, Zachariou M, Sokratous K, Bourdakou MM, Spyrou GM. Systems bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches. Briefings Bioinform 2017.

[3] Domingo-Fernandez D, Mubeen S, Marin-Llao J, Hoyt CT, Hofmann-Apitius M. PathMe: merging and exploring mechanistic pathway knowledge. BMC Bioinform 2019;20:243.

[4] Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 2003;31:258–61.

[5] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90–7.

[6] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res 2007;35:W186–92.

[7] Gerstner N, Kehl T, Lenhof K, Müller A, Mayer C, Eckhart L, et al. GeneTrail 3: advanced high-throughput enrichment analysis. Nucleic Acids Res 2020;48:W515–20.

[8] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8.

[9] Kanehisa, M. (2002) The KEGG database. Novartis Foundation symposium, 247, 91–101; discussion 101–103, 119–128, 244–152.

[10] Fabregat A, Sidiropoulos K, Viteri G, Marin-Garcia P, Ping P, Stein L, et al. Reactome diagram viewer: data structures and strategies to boost performance. Bioinformatics 2018;34:1208–14.

[11] Kelder T, Van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res 2011;40:D1301–7.

[12] Qiu Y, Yu-Qing Q. KEGG pathway database. Encyclop Syst Biol 2013:1068–9.

[13] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2018;46:D649–55.

[14] West DB. Introduction to Graph Theory. Upper Saddle River: Prentice Hall; 2001.

[15] Csardi MG. Package 'igraph'. Last accessed, 3, 2013 2013.

[16] Minadakis G, Zachariou M, Oulas A, Spyrou GM. PathwayConnector: finding complementary pathways to enhance functional analysis. Bioinformatics 2019;35:889–91.

[17] Minadakis G, Spyrou GM. Computational Methods in Synthetic Biology. Springer; 2021. p. 231–49.

[18] Kakouri AC, Votsi C, Tomazou M, Minadakis G, Karatzas E, Christodoulou K, et al. Analyzing gene expression profiles from ataxia and spasticity phenotypes to reveal spastic ataxia related pathways. Int J Mol Sci 2020;21:6722.

[19] Karatzas, E., Zachariou, M., Bourdakou, M., Minadakis, G., Oulas, A., Kolios, G., et al., PathWalks: Identifying pathway communities using a disease-related map of integrated information. bioRxiv, 2020.

[20] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90–7.

[21] Prestori F, Moccia F, D'Angelo E. Disrupted calcium signaling in animal models of human spinocerebellar ataxia (SCA). Int J Mol Sci 2020;21:216.

[22] Takashima H, Boerkoel CF, John J, Saifi GM, Salih MA, Armstrong D, et al. Mutation of TDP1, encoding a topoisomerase I–dependent DNA damage repair enzyme, in spinocerebellar ataxia with axonal neuropathy. Nat Genet 2002;32:267–72.

[23] Kostrzewa M, Klockgether T, Damian MS, Müller U. Locus heterogeneity in Friedreich ataxia. Neurogenetics 1997;1:43–7.

[24] Nambo-Venegas R, Valdez-Vargas C, Cisneros B, Palacios-González B, Vela-Amieva M, Ibarra-González I, et al. Altered plasma acylcarnitines and amino acids profile in spinocerebellar ataxia type 7. Biomolecules 2020;10:390.

[25] Barbeau A. Friedreich's Ataxia 1980 an overview of the physiopathology. Can J Neurol Sci 1980;7:455–68.

[26] Dahl ES, Aird KM. Ataxia-telangiectasia mutated modulation of carbon metabolism in cancer. Front Oncol 2017;7:291.

[27] Purkiss P, Baraitser M, Borud O, Chalmers R. Biochemical and clinical studies of Friedreich's ataxia. J Neurol Neurosurg Psychiatry 1981;44:574–80.

[28] Surtees R. Understanding neurodegenerative disorders. Curr Paediatr 2002;12:191–8.

[29] Ronchi D, Monfrini E, Bonato S, Mancinelli V, Cinnante C, Salani S, et al. Dystonia-ataxia syndrome with permanent torsional nystagmus caused by ECHS1 deficiency. Ann Clin Transl Neurol 2020;7:839–45.

[30] Bhandari, J., Thada, P.K. and Samanta, D. (2020), StatPearls [Internet]. StatPearls Publishing.

[31] Zachariou M, Minadakis G, Oulas A, Afxenti S, Spyrou GM. Integrating multi-source information on a single network to detect disease-related clusters of molecular mechanisms. J Proteom 2018;188:15–29.

[32] Kakouri AC, Christodoulou CC, Zachariou M, Oulas A, Minadakis G, Demetriou CA, et al. Revealing clusters of connected pathways through multisource data integration in huntington's disease and spastic ataxia. IEEE J Biomed Health Inform 2018;23:26–37.

[33] Gkretsi V, Louca M, Stylianou A, Minadakis G, Spyrou G, Stylianopoulos T. Inhibition of breast cancer cell invasion by ras suppressor-1 (RSU-1) silencing is reversed by growth differentiation factor-15 (GDF-15). Int J Mol Sci 2019;20:163.

[34] Lu X, Xue L, Sun W, Ye J, Zhu Z, Mei H. Identification of key pathogenic genes of sepsis based on the Gene Expression Omnibus database. Mol Med Rep 2018;17:3042–54.

[35] Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330.

[36] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009;25:1091–3.

[37] Minadakis G, Fuentes AM-P, Tsouloupas G, Papatheodorou I, Spyrou GM. PathExNET: a tool for extracting pathway expression networks from gene expression statistics. Comput Struct Biotechnol J 2021;19:4336–44.

[38] Lemsara A, Ouadfel S, Fröhlich H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. BMC Bioinform 2020;21:1–20.

[39] Palombo V, Milanesi M, Sferra G, Capomaccio S, Sgorlon S, D'Andrea M. PANEV: an R package for a pathway-based network visualization. BMC Bioinform 2020;21:1–7.

[40] Domingo-Fernández D, Hoyt CT, Bobis-Álvarez C, Marín-Llaó J, Hofmann-Apitius M. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. NPJ Syst Biol Appl 2018;4:1–8.

[41] Pita-Juárez Y, Altschuler G, Kariotis S, Wei W, Koler K, Green C, et al. The pathway coexpression network: revealing pathway relationships. PLoS Comput Biol 2018;14:e1006042.

[42] Kohl M, Wiese S, Warscheid B. Data mining in proteomics. Springer; 2011. p. 291–303.

[43] Emmert-Streib F, Dehmer M. Networks for systems biology: conceptual connection of data and function. IET Syst Biol 2011;5:185–207.

[44] Najafi A, Bidkhori G, Bozorgmehr JH, Koch I, Masoudi-Nejad A. Genome scale modeling in systems biology: algorithms and resources. Curr Genom 2014;15:130–59.