PathJam: a new service for integrating biological pathway information

Daniel Glez-Peña[1], Miguel Reboiro-Jato[1], Rubén Domínguez[2], Gonzalo Gómez-López[3], David G. Pisano[3], Florentino Fdez-Riverola[1*]

[1]Escuela Superior de Ingeniería Informática University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain

[2]Informatics Unit, Complexo Hospitalario de Ourense, C/ Ramón Puga 54, 32005, Ourense, Spain

[3]Unidad de Bioinformática (UBio), Programa de Biología Estructural y Biocomputación, Centro Nacional de Investigaciones Oncológicas (CNIO), C/ Melchor Fernández Almagro 3, 28029, Madrid, Spain

{dgpena,mrjato,riverola}@uvigo.es

ruben.dominguez.carbajales@sergas.es

{ggomez,dgonzalez}@cnio.es

Summary

Biological pathways are crucial to much of the scientific research today including the study of specific biological processes related with human diseases. PathJam is a new comprehensive and freely accessible web-server application integrating scattered human pathway annotation from several public sources. The tool has been designed for both (*i*) being intuitive for wet-lab users providing statistical enrichment analysis of pathway annotations and (*ii*) giving support to the development of new integrative pathway applications. PathJam's unique features and advantages include interactive graphs linking pathways and genes of interest, downloadable results in fully compatible formats, GSEA compatible output files and a standardized RESTful API.

1 Introduction

Over the last years, the research community is experiencing massive growth in biological data, from genome and metagenome sequencing to high-throughput assays and microarray studies. In this situation, attempting to understand individual genes on a list of significant genes is demanding and laborious. In the middle of the exponential growth of diverse types of biological knowledge, pathways are critical to understanding the functions of individual genes and proteins in terms of systems and processes that contribute to normal physiology and to disease [1]. In this context, biological pathways present a special case in which the information is not directly coupled to data collection. As a result, pathway information is particularly challenging to compile, standardize and curate, keeping it as accessible, up-to-date, and integrated as possible [2].

In order to deal with the vast amount of available representation formats in system biology, advances in the definition of standardized formats for representing both data and interactions

* Corresponding author

have been developed during last years [3]. Within the context of biological pathways, SBML (*Systems Biology Markup Language*) [4], CellML [5], PSI MI (*Proteomics Standards Initiative's Molecular Interaction*) [6] and BioPAX (*Biological Pathway Exchange*) [7] representation languages and ontologies have emerged as possible candidates to guide semantic integration in the future [8]. In an interrelated way, it has also been challenging to integrate and make accessible all the plethora of existing pathway-related information and resources. A good starting point is Pathguide [9], which provides a single access point for biological pathway and network databases.

The aforementioned explosion of high-throughput data, together with the publicly available databases containing valuable pathway-related information, have generated unprecedented opportunities for investigation of concerted changes that disrupt biological functions. This situation has motivated an increasingly need of new computational tools for integrated pathway analysis. These tools, capable of analyzing high-throughput data inside a biological pathway context, range from commercial software packages including Metacore [10], Ingenuity Pathways Analysis [11], PathwayAssist [12] or PathArt [13] to an acceptable number of publicly accessible applications such those summarized in Table 1.

Each of these tools has unique features that distinguish it from others covering related functionalities as (*i*) displaying microarray data within the context of annotated pathways, (*ii*) providing statistical assessment of the reliability of differentially expressed genes, (*iii*) visualizing and analyzing gene networks using an interactive GUI or (*iv*) accessing curated biological pathway/association network databases, among others. Despite all these software tools available for giving support in different phases of an integrative analysis, the complex interaction between pathways and the involvement of pathways in multiple phenotypes notably complicate the frequent and basic task of interpreting and comparing gene expression patterns.

In the wet-lab scenario, biomedical researchers frequently want to quickly browse relationships between pathogenesis and biological pathways regulated by differentially expressed gene lists [31]. In this situation, some of the limitations of available tools reported in Table 1 are (*i*) the inadequacy to handle simple gene lists in an intuitive way, (*ii*) the inability of easily perform pathway enrichment analysis as well as their clear interpretation and (*iii*) the impossibility of exporting obtained data to another gene set enrichment analysis applications in order to carry out subsequent testing [32]. This situation also resonates with the large portion of the biology community mainly interested in basic and straightforward statistical pathway analysis and figures for publications and presentations.

From a developer perspective, in a field where integration is of utmost importance, open-source software is a necessity to further encourage cooperative development guarantying at the same time the quality of available bioinformatics tools. Up to now, except for DGEM software [16] currently not accessible, GenMAPP application [18, 19] in which multiplatform is not supported, KOBAS server [20] only giving access to KEGG pathways and PathVisio editor [24] lacking of statistical support, there is no a reference application with source code available for integrating pathway databases. Moreover, none of the previous tools implement an accessible API able to provide the programmatic facilities needed by specialized bioinformaticians for unifying the access to biological pathway data.

To facilitate both (*i*) the execution of common analysis of large gene lists based on pathway-related information as well as their subsequent interpretation and refinement using specific gene set enrichment tools and (*ii*) the cooperative development of more sophisticated applications based on a fully open source schema, we have developed an on-line publicly available service called PathJam.

	Licensing	**Multiplatform**	**Input**		**Process**		**Output**	**Developer issues**	
Tool			*Login required?*	*Support for multiple namespaces*	*Supported Pathway Databases*	*Statistical testing support?*	*Exporting capabilities*	*Source code available?*	*Interoperability with other tools?*
ArrayXPath [14,15]	Not defined	Yes (Web-based)	Yes (to access some functionalities)	Yes	4 GenMAPP PharmGKB KEGG BioCarta	Yes (Fisher's exact test)	No	No	No
DGEM [16] *Not accessible*	Not defined	Yes (Web-based)	NA	No	1 KEGG	Yes (t-test with p-values adjusted using FDR)	No	Yes (freely available on request to the authors)	No
Eu.Gene [17]	Not defined (free of charge for academic users)	Yes Stand-alone (Java-based)	No	Yes	3 GenMAPP KEGG Reactome	Yes (Fisher's exact test & GSEA)	Yes (HTML & MS-Excel)	No	No
GenMAPP [18,19]	Open-source (Apache)	No (Windows OS) Stand-alone Visual Basic 6.0 application	Yes, mandatory	Yes	4 GenMAPP *curated*: KEGG BioCarta PharmGKB	Yes (permutation test with p-values adjusted using Westfall-Young)	Yes (HTML)	Yes	Yes (MAPPFinder & MAPPBuilder from the same software package)
KOBAS [20]	Biopython license	Yes (Web-based &	Yes (to access some	Yes	1 KEGG	Yes (Binomial, χ^2, Fishers' exact	No	Yes	No

	Licensing	**Multiplatform**	**Input**		**Process**		**Output**	**Developer issues**	
Tool			*Login required?*	*Support for multiple namespaces*	*Supported Pathway Databases*	*Statistical testing support?*	*Exporting capabilities*	*Source code available?*	*Interoperability with other tools?*
		Stand-alone)	functionalities)			& Hypergeom. Adjusted using FDR)			
PathExpress [21, 22]	Not defined	Yes (Web-based)	No	Only two: Affymetrix or Gene Accession Numbers	1 KEGG	Yes (p-values adjusted using Bonferroni or FDR)	Yes (graphics as PNG & data as tab-delimited text files)	No	No
PathMAPA [23] *No longer available*	Not defined	Yes (Web-based)	NA	Only a expression data file	3 KEGG TAIR NCBI	Yes (Fisher's exact test with adjusted p-values)	Yes (resulting graph)	No	No
PathVisio [24]	Open-source (Apache v2.0)	Yes Stand-alone (Java-based)	No	Yes	Those supported by GenMAPP	Yes	Yes (using Batik SVG toolkit for graphics)	Yes (from a SVN repository	Yes (applet version used by WikiPathways)
Pathway Explorer [25]	Not defined (free of charge for academic users)	Yes (Web-based & Stand-alone)	Yes (in web-based for accessing some functionalities)	Yes	3 KEGG BioCarta GenMAPP	Yes (Fisher's exact test with adjusted p-values)	Yes (graphics as PNG, JPG and SVG & data as tab-delimited text files)	No	No
Pathway Express [26]	Not defined	Yes (Web-based)	Yes, mandatory	Yes	1 KEGG	Yes (using a	Yes	No	Yes (Onto-Tools &

	Licensing	**Multiplatform**	**Input**		**Process**		**Output**	**Developer issues**	
Tool			*Login required?*	*Support for multiple namespaces*	*Supported Pathway Databases*	*Statistical testing support?*	*Exporting capabilities*	*Source code available?*	*Interoperability with other tools?*
						biologically-inspired *impact factor* index)			Bioconductor)
Pathway Miner [27]	Not defined	Yes (Web-based)	Yes, mandatory	No	3 KEGG BioCarta GenMAPP	Yes (Fisher's exact test)	No	No	No
Pathway Processor [28] *Not accessible*	Not defined	Yes Stand-alone (Java-based)	No	No	1 KEGG	Yes (Fisher's exact test)	Yes (data as tab-delimited text files)	No	No
VisANT [29, 30]	Not defined	Yes (Web-based & Stand-alone Java-based)	Yes (to access some functionalities)	Yes	1 KEGG	Yes (Hypergeom. test-based algorithm)	Yes (graphics as JPG and SVG & data as tab-delimited text files)	No	No
PathJam	GNU LGPL (Lesser General Public License)	Yes (Web-based)	No	Yes	4 KEGG NCI BioCarta Reactome	Yes (Fisher's exact test with adjusted p-values)	Yes (graphics as PNG & cross-tables as CSV)	Yes (direct download)	Yes (full integration with CARGO)

Table 1: Comparison of PathJam with existing tools.

PathJam is able to gather biological pathway data coming from various sources of information, integrate them, and provide key features for supporting functional study of gene lists of interests. Currently, PathJam supports human pathways from Reactome [33] KEGG [34], NCI [35] and BioCarta [36] databases. PathJam users can specify their gene lists as input by using any of the supported identifiers: (*i*) Affymetrix probesets, (*ii*) Entrez gene IDs, (*iii*) Ensembl IDs or (*iv*) Swiss-Prot identifiers. Our deployed application not only supports many unique ways to query, integrate and analyze pathway data, but also combines advantages of visualizing it with a more intuitive pathway network-based representation, and many other features that allow for more comprehensive data acquisition and analysis.

2 Methods

2.1 Technical architecture

PathJam is implemented as a client/server application in the Java programming language. It can be accessed via its web site (http://www.pathjam.org), which contains an embedded Java applet. This component transparently runs in the client machine, by performing regular underlying HTTP requests to gather server information. In addition, our server also implements a RESTful (*Representational State Transfer*) web service API suitable for programmers who want to use PathJam indexed information in their own developments. The server runs on Apache Tomcat 5 and the applet requires the Sun Java browser plugin version 5 or later. PathJam has been successfully tested in Internet Explorer 7, Firefox 3, Opera 9.62 and Safari 3 browsers working on Windows XP/Vista, Ubuntu Linux 8.04 version and Mac OSX 10.5 of Intel architecture.

Up to now, PathJam integrates definitions of pathways from 4 different sources: KEGG, Reactome, NCI Pathway Interaction Database (NCI-PID) and Biocarta. Each database makes access to its contents through one or more up-to-date technologies. We gather information from KEGG using its SOAP-based web service; Reactome is available through the BioMart REST-based web service; NCI-PID and Biocarta data are retrieved by directly accessing the NCI-PID website (using HTML parsing and resource downloads).

From each datasource, PathJam only obtains (*i*) a list of pathways (name and database identifier) and (*ii*) a list of gene identifiers of each pathway. Table 2 summarizes how this information is obtained from the external resources.

Resource	Technology	Pathway list	Gene list	Namespace
KEGG	KEGG WS API	'list_pathways' function	'get_genes_by_pathway' function	Entrez
Reactome	BioMart	XML Query 1	XML Query 2	Entrez
NCI-PID	HTML/FTP to NCI-PID website	Parsing the 'browse pathways' section in the site	Molecule list export facility through FTP (csv file). ftp://ftp1.nci.nih.gov/pub/PID /molList/{pathid}.mol.csv.gz	Entrez
Biocarta	HTML/FTP to NCI-PID website			

Table 2: External sources and accession methods used by PathJam for integrating pathway information.

2.2 Data model

In order to integrate and actualize pathway data from the external supported databases, PathJam server manages an internal 'index' that is periodically updated. The objective of this index is to maintain existing, and discover new, pathway-gene associations. However, we should take into account that pathway sources may not use the same namespace to identify genes/proteins. It is not the case of the current supported resources (all offering Entrez-gene IDs), but it is crucial for future developments to being able to integrate pathways whose entries are specified in different namespaces. In addition, PathJam application needs to recognize common gene names as input, such as HGNC-Symbols (gene names), Entrez-Gene, Ensembl, Uniprot/Swissprot and Affymetrix probeset IDs. For example, if the user queries 'BRCA2' (HGNC-Symbol) or 'ENSG00000139618' (Ensembl ID), PathJam has to return all pathways found in available sources where the Entrez-gene ID '675' is involved. In order to overcome this circumstance, the index is generated in two phases.

First, and previously to gather any external information, we build a local 'dictionary' structure based on Ensembl Gene database (version 55) which can be accessed via its BioMart Service. Our dictionary keeps all protein-coding genes uniquely identified by their Ensembl ID together with all their possible names across different namespaces. In such a situation, it is common to find multiple interrelations, that is, one Ensembl identifier corresponding to several Entrez, Uniprot, HGNC (rare cases) and/or probeset identifiers (whilst the inverse case is very infrequent). Table 3 takes a snapshot of the dictionary contents and the relationships between their namespaces.

Namespace	**Total entries**	**Ensembl-Other (*Ensembl IDs* with several *identifiers* in the namespace)**	**Other-Ensembl (*Identifiers* in the namespace with several *Ensembl IDs*)**
Ensembl	21416	-	-
HGNC	18864	221 (1%)	0 (0%)
Entrez	19831	8551 (39%)	632 (3%)
Uniprot/Swissprot	19351	9389 (43%)	319 (1%)
Affymetrix Probeset	30854	10284 (48%)	1513 (7%)

Table 3: PathJam dictionary summary. Built from Ensembl 55 Homo sapiens database (22/09/2009).

In the second phase, each source database is accessed and their genes, given in a particular namespace, are fetched from the dictionary and linked with corresponding pathways. Each gene is indexed as a so called 'virtual gene', corresponding to one entire row of the dictionary (or more than one row, in rare cases), so it can be found regardless of the namespace being used. Figure 1 represents the whole process (steps one and two) in a schematic way.

2.3 Standardized RESTful Interface for exporting functionalities

The PathJam web service API allows programmers to access the PathJam index in a programmatic way. Using this service, the server can be accessed from multiple programming languages, allowing researchers to wire PathJam results to their experiments. Currently, the API offers a pathway retrieval service given a list of gene identifiers, a gene retrieval service given a list of pathways, and some other functionality. Table 4 summarizes the available options in the API.

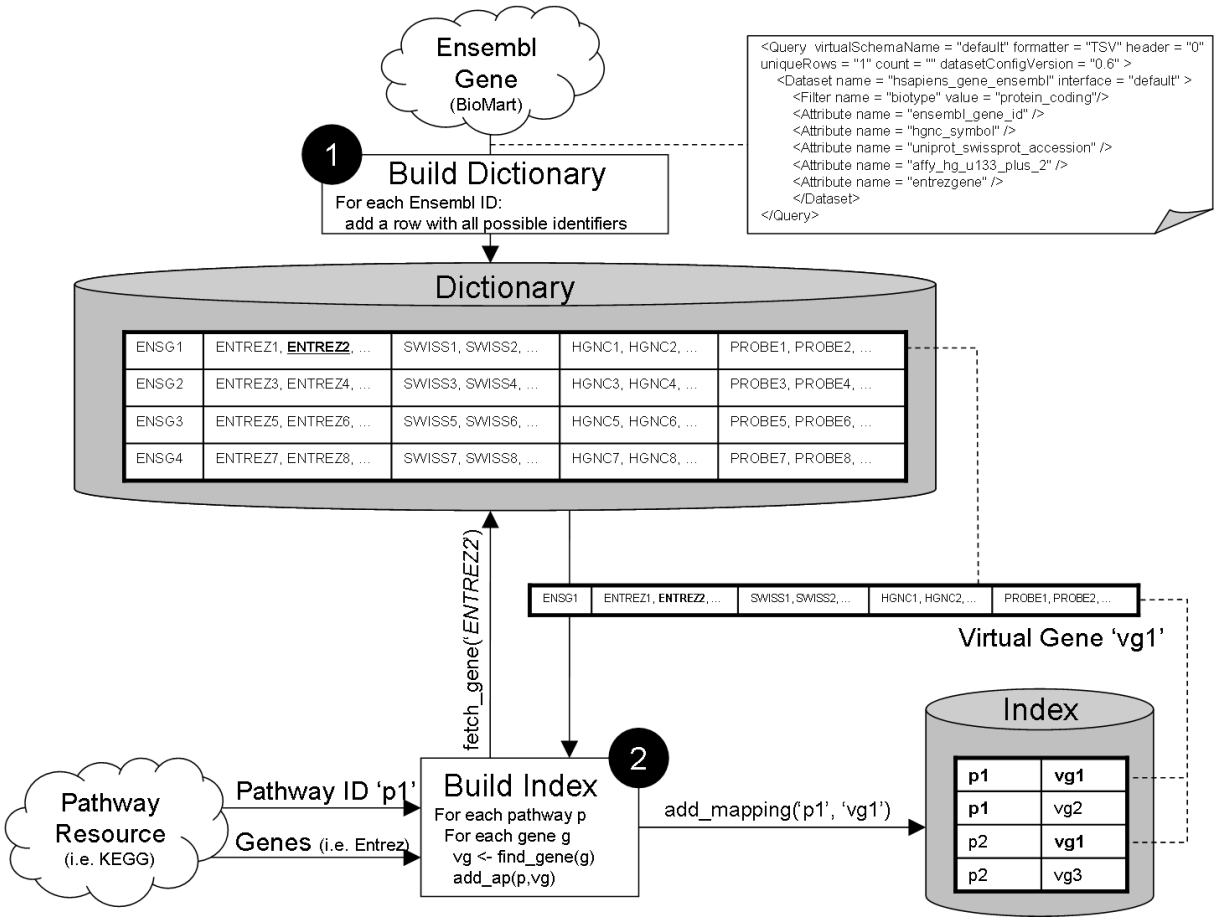

Figure 1: PathJam Dictionary containing virtual genes. Different namespaces are integrated through the aggregation of different identifiers belonging to the same virtual gene.

Function	**Description**
Source Listing	Lists all data sources currently indexed
	Example URLs http://www.pathjam.org/server/api/sources
Pathway Listing	Lists all pathways currently indexed
	Example URLs http://www.pathjam.org/server/api/pathways http://www.pathjam.org/server/api/pathways/reactome
Pathways for Genes	Retrieves the pathways from a list of gene identifiers belonging to any namespace. It returns one line per gene containing all the pathways where each gene is involved
	Example URLs http://www.pathjam.org/server/api/pathways?genes=BRCA1,TP53
Genes for Pathways	Retrieves the genes from a list of (*i*) pathway identifiers (previously obtained with the pathway listing function) or (*ii*) pathway names (without white spaces). It returns one line per pathway containing all the genes involved in each pathway
	Example URLs http://www.pathjam.org/server/api/genes?pathways=path:hsa00010,REACT_1707 http://www.pathjam.org/server/api/genes?pathways=pyrimidinemetabolism

Table 4: Core functionality provided by PathJam RESTful API.

Like any RESTful web service, operations are performed via web queries with a well-defined URL structure. The PathJam server API is located at http://www.pathjam.org/server/api/

3 Results and discussion

3.1 Overview

In recent years, some valuable wet-lab oriented tools have been developed in order to facilitate the functional analysis of gene lists within the biological context of molecular pathways. In this scenario, previous successful applications are DAVID software [37] and FatiGO+ server [38], which provide useful information to understand gene lists.

In contrast to other functional analysis tools, PathJam server application integrates pathway-related annotations coming from distributed public sources (like Reactome, KEGG, BioCarta and National Cancer Institute) displaying pathway annotations and functional enrichment results simultaneously. In this way, the functional analysis results from a number of different pathway annotation sources are analyzed together, therefore facilitating the biological understanding of gene lists of interest. In order to produce interpretable results, PathJam implements a two-tailed Fisher's exact test where adjusted p-values (q-values) are obtained by Bonferroni multiple testing correction [39]. Additionally, the server can also answer questions such as: a) in which pathways is implicated a user-selected gene (or gene set)? and b) which genes are involved in a given pathway (or set of pathways)?. Furthermore, PathJam has been designed to generate compatible outputs with well-known gene set-based methods such as GSEA gmx file format [40]. All these features together with the standardized RESTful interface implemented in PathJam are not currently available from any other functional analysis server.

Therefore, PathJam users can download and customize gene sets related to genes of interest in order to use them in gene set enrichment analyses (i.e. all pathways in which a given gene is implicated). Moreover, PathJam users can obtain interactive graphs linking genes with pathway annotations as well as exportable result tables and figures in fully compatible file formats (.txt, .xls, .png) with popular applications in the wet-labs (i.e. MS Office, Open Office, etc.).

3.2 PathJam graphical user interface

The user interaction with the PathJam server starts by performing a simple query containing a list of genes of interest introduced via a web form. Gene names can be specified using any of the supported namespaces: Ensembl, HGNC Symbol, Entrez, Uniprot/Swissprot accession or Affymetrix probeset. Once the list is submitted, a Java applet is locally started in the client browser and the main interface is displayed.

The PathJam applet is organized in two main zones. Figure 2 shows the aspect of PathJam when the user has submitted his query.

On the left side, the user can see related information about the submitted gene list and all pathways in which those genes are involved. The right side is intended to give support to several interactive views and is also structured in two different tabs: 'Pathways for selected genes' and 'Genes for selected pathways'. These two tabs are enabled if some genes or some pathways are selected in the lists showed on the left zone, respectively.

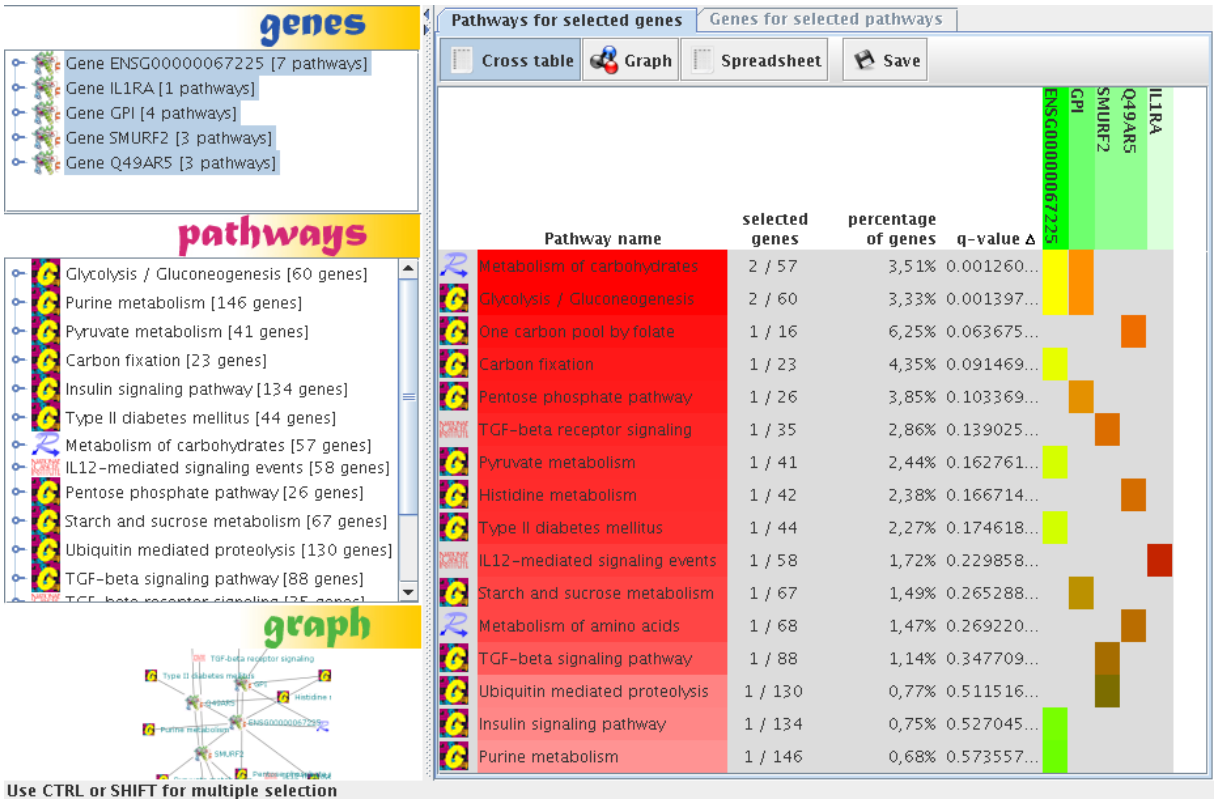

Figure 2: PathJam screenshot showing the results obtained from a simple list of genes. A two-tailed fisher's exact test gives information about enriched biological pathways with adjusted p-values using multiple testing correlation. Columns can be ordered by any criterion..

In the 'Pathways for selected genes' tab, the first available view is the 'Cross table', containing the user submitted genes (arranged in columns) and the pathways including, at least, one of those genes (arranged in rows). This view also implements a two-tailed Fisher's exact test in order to identify enriched biological pathways. By default, pathways are ranked and displayed in red tones (following the score obtained in the test) whilst genes are sorted and coloured in green tones (following the number of pathways in which they are involved). The user can interactively modify this predefined organization by clicking in the column titles, being also possible to sort the data by more than one column simultaneously.

Each cell in the matrix (pathways x genes) graphically represents the intersection of a pathway p_i and a gene g_j. The colour assigned to each cell, c_{ij}, is gray if g_j is not present in p_i, or a combination of the p_i colour (red tone) and the g_i colour (green tone) if g_j is involved in p_i.

A second useful view is the 'Graph' representation (see Figure 3). This component implements an interactive, exportable, automatic-layout visual graph for linking genes and pathways. This representation is available for both 'Pathways for selected genes' and 'Genes for selected pathways' tabs. It allows the user to see an overview of the gene-pathway relationships in an interactive way by using a physical force simulation engine for dynamic layout and animation. The 'Graph' component of PathJam application is implemented making use of the prefuse toolkit [41], a publicly available Physics engine that uses the Java 2D graphics library for information visualization.

A third component provides a raw-data view named 'Spreadsheet', which is also available in the two tabs. The spreadsheet contains genes or pathways as column titles and their associated pathways or genes as row lines, depending on the current active tab: 'Pathways for selected genes' or 'Genes for selected pathways', respectively. This representation is suitable for

exporting the results obtained from the PathJam index via copy/paste in order to perform further analysis. Figure 4 illustrates the implemented spreadsheet view.

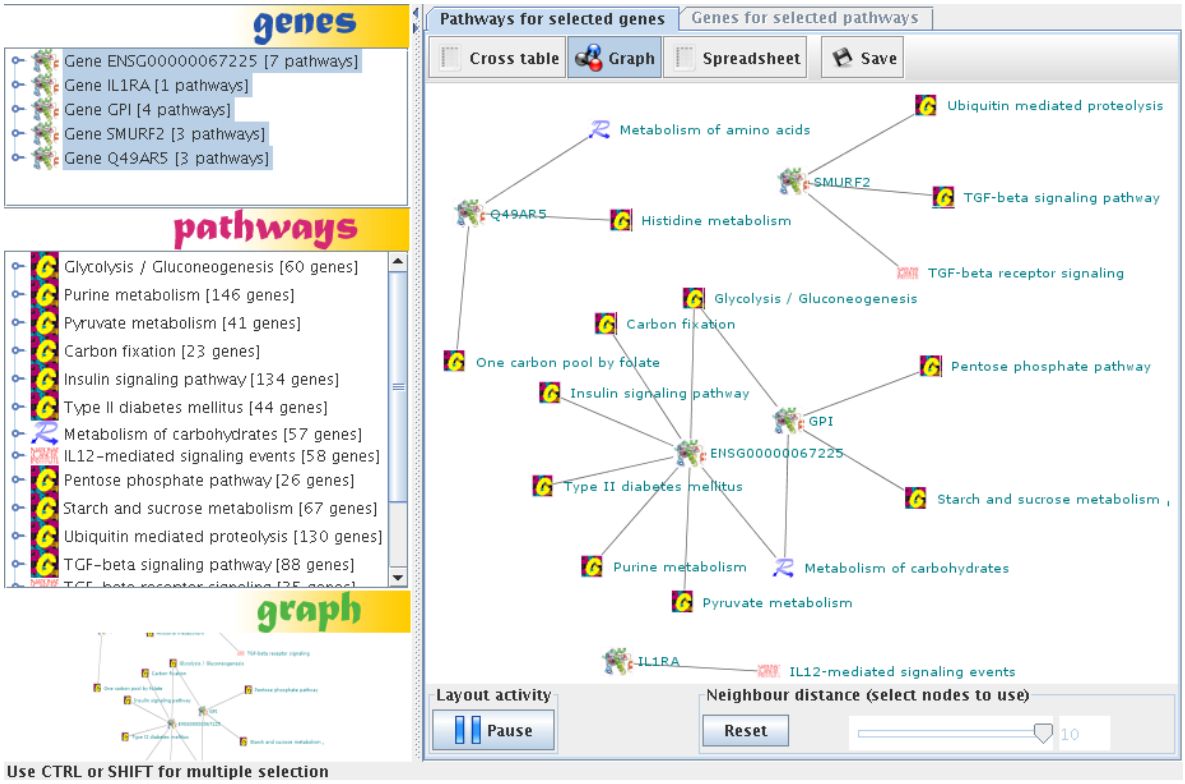

Figure 3: Interactive graph component representing identified interactions between genes and pathways. The graphical representation can be customized by the user hiding unwanted associations (based on a neighbour distance principle). All the graphics elements (genes and pathways) are clickable for obtaining specific information.

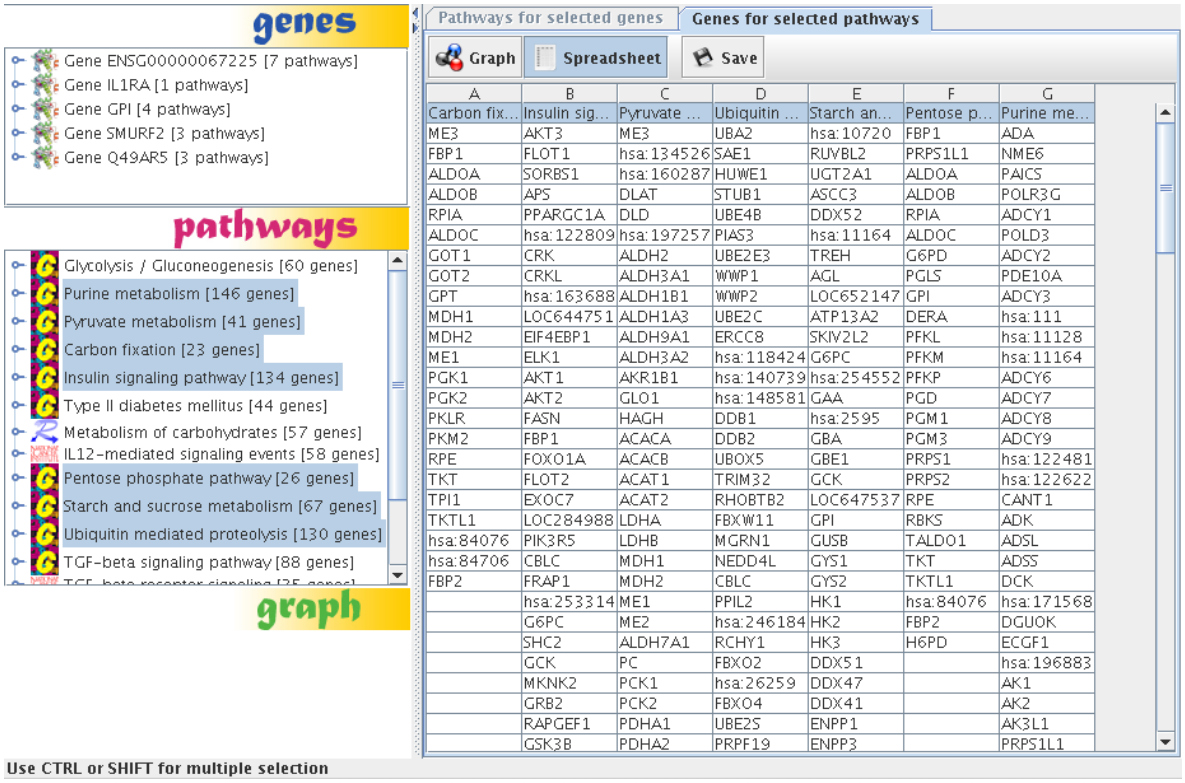

Figure 4: Exportable raw-data view of PathJam results. The spreadsheet layout of PathJam allows users to easily copy/paste obtained results to other applications.

Moreover, each active view can be exported to a suitable format: 'Cross table' and 'Spreadsheet' can be saved to a .csv text file and the 'Graph' can be saved to a .png file. In addition, if the 'Spreadsheet' is showing the 'Genes for selected pathways', it is also possible to generate a GSEA compatible output file (.gmx) containing user-defined collections of gene sets related to pathways of interest.

3.3 PathJam widget for CARGO

From a different perspective, a new innovative approach named CARGO (*Cancer And Related Genes Online server*) has recently been released [42]. CARGO has been designed with the aim of integrating disperse information stored in biological databases. Thus, CARGO has been implemented using configurable web widgets to display and visualize valuable information of genes from scattered biological databases independently of their native format or access method. From this perspective, CARGO represents a novel and intuitive way to retrieve and integrate information for a given gene of interest. In this work, we have also developed and made accessible a single-gene query version of PathJam for exporting part of the available functionality to the CARGO server. Figure 5 shows a related screenshot.

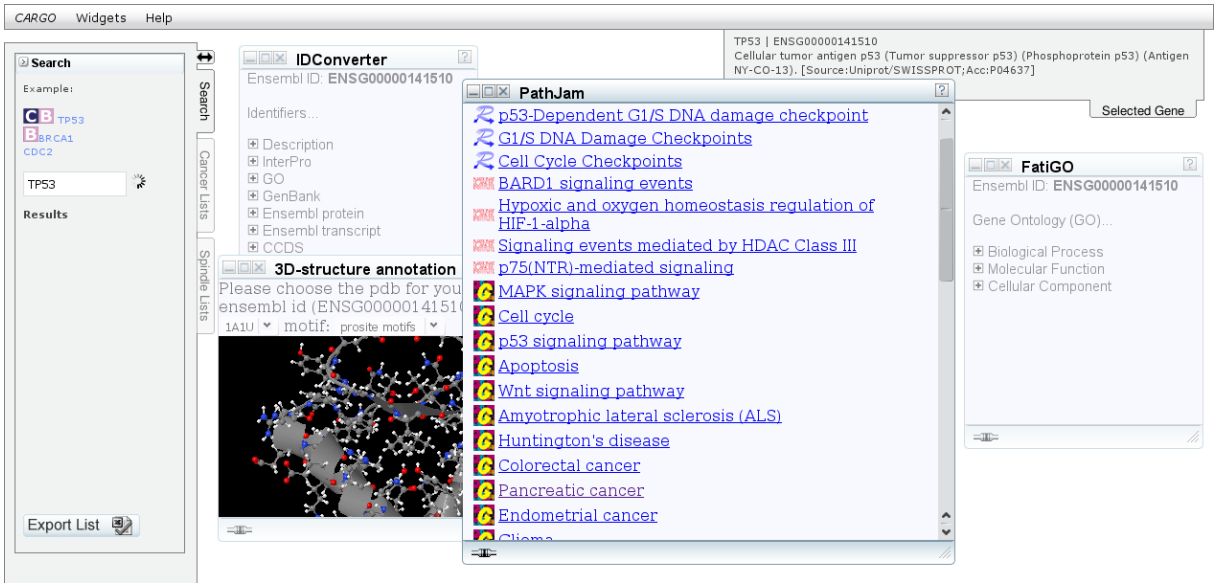

Figure 5: PathJam Widget for CARGO users showing information related with TP53 gene.

3.4 Key features

The main contribution of PathJam server is linking gene expression data within a pathway context and their application to gene set enrichment analysis in an understandable way. PathJam application is freely available under an open source license providing an easily access to unique functionalities not found in a single similar application. The key features provided by PathJam are detailed below.

3.4.1 Flexible pathway data gathering and integration of multiple namespaces

Due to the great number and the diversity of available database resources continuously publishing pathway related data, it is very difficult to maintain updated information about existing relations between genes and pathways. To solve this problem, PathJam relies in an internal periodically updated indexing structure for integrating and actualizing pathway data coming from external sources. Moreover, to cope with the recurrent biological problem of

identical data described by different databases using multiple identifiers, PathJam constructs and manages a unique local dictionary where the concept of 'virtual gene' is used in order to maintain accessible all protein-coding genes through their Ensembl ID.

3.4.2 Intuitive web interface for querying and analyzing large gene lists of interests

PathJam supports user queries of individual genes or gene lists, which can be specified by simultaneously using multiple identifiers, via a standard web form. Once the server receives the user query, an embedded Java applet is transparently loaded in the client machine for maintaining the current active session with the user. The PathJam applet provides an intuitive interface for supporting user interaction in an interactive way by performing regular underlying HTTP requests to gather server information. Multiple views of pathway-related information are available showing integrative results about the user query.

3.4.3 Standardized web service for exporting functionalities

Data indexed by PathJam is accessible for query and export via a standard-based web service interface. The PathJam web service is not dependent on a specific operating system or programming language, and uses a REST-based architecture. By implementing this feature, PathJam enables interoperable communication with other software modules, and enables third-party applications to more easily build and expand tools for visualization, analysis and model simulation. This helps lower the development effort required to interface with PathJam, while simultaneously maintaining platform and language independence.

3.4.4 GSEA interoperability

In addition to the functionalities implemented for achieving an intuitive and user-friendly framework for biological pathway analysis of human gene lists, PathJam has been designed to produce compatible outputs with well-established gene set enrichment methods like GSEA [40]. This unique characteristic allows the user to generate compatible gmx (Gene MatriX) files describing newly assembled gene sets for subsequent functional analysis. As PathJam evolves, future developments will directly support more input formats for available bioinformatics enrichment tools [32].

4 Conclusions

With the goal of facilitating the pathway analysis of gene lists generated by high-throughput experiments we have introduced PathJam. The main objective of our application is to guarantee the simplicity of use for both experimental biologists and specialized bioinformaticians who need assistance in biological pathway analysis. The application is an intuitive and freely accessible server integrating pathway information available in scattered public repositories. Given a list of genes (or proteins) of interest the server provides several utilities to final users like (*i*) an integrated pathway annotation (KEGG, Reactome, NCI-PID and BioCarta) for each element of a given gene list; (*ii*) access to detailed information in public resources about pathways and genes (links to GeneCards); (*iii*) pathway enrichment analysis including FDR adjustment to account for multiple testing; (*iv*) interactive 2D graphs linking pathways and genes of interest; (*v*) downloadable results including tables and figures in fully compatible formats with popular software in wet-labs; (*vi*) GSEA compatible output files (.gmx) containing user-defined collections of gene sets related to pathways of interest; and (*vii*) full integration with CARGO through PathJam widget.

The PathJam server has been running for a period of nine months. During this time, three independent beta testers have identified and reported weak points and specific issues which were corrected prior to the release of the actual stable version of PathJam.

Acknowledgements

We thank JM Rodríguez and the National Institute of Bioinformatics (www.inab.org), a platform of Genoma España for providing useful information and advice in CARGO widget implementation. M Sánchez-Beato and MA Piris for allowing the use of Tracey et al. gene list for PathJam validation. This work is supported in part by the projects *Development of biomedical applications* from University of Vigo (Spain) and *MEDICAL-BENCH: Platform for the development and integration of knowledge-based data mining techniques and their application to the clinical domain* (TIN2009-14057-C03-02) from Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government and the European Union. The work of D Glez-Peña is supported by an Angeles Alvariño contract from Xunta de Galicia. G Gómez-López is funded in part by the Spanish National Institute of Bioinformatics (INB), a platform of Genoma España.

References

[1] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin and C. Evelo. WikiPathways: Pathway Editing for the People. PLoS Biol, 6(7):e184, 2008.

[2] H. N. Kasamsetty, X. Wu and J. Y. Chen. Towards an integrative human pathway database for systems biology applications. In Proceedings of the 2008 ACM Symposium on Applied Computing: 16-20 March 2008, Fortaleza, Ceara, Brazil. ACM New York: NY, USA, 1297-1301, 2008.

[3] L. Strömbäck, D. Hall and P. Lambrix. A review of standards for data exchange within systems biology. Proteomics, 7(6):857-867, 2007.

[4] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner and J. Wang. SBML Forum: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics, 19(4):524-31, 2003.

[5] C. M. Lloyd, M. D. Halstead and P. F. Nielsen. CellML: its future, present and past. Prog Biophys Mol Biol, 85(2-3):433-50, 2004.

[6] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue and R. Apweiler. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol, 22(2):177-183, 2004.

[7] BioPAX, Biological Pathways Exchange (http://www.biopax.org/).

[8] J. S. Luciano and R. D. Stevens. e-Science and biological pathway semantics. BMC Bioinformatics, 8(3):S3, 2008.

[9] G. D. Bader, M. P. Cary and C. Sander. Pathguide: a Pathway Resource List. Nucleic Acids Res, 34(Database issue):D504-D506, 2006.

[10] MetaCore, a product of GeneGO Inc (http://www.genego.com).

[11] Ingenuity Pathways Analysis tool, a product of Ingenuity Systems Inc (http://www.ingenuity.com).

[12] A. Nikitin, S. Egorov, N. Daraselia and L. Mazo. Pathway studio – the analysis and navigation of molecular networks. Bioinformatics, 19:1-3, 2003.

[13] PathArt, a product of Jubilant Biosys Ltd (http://www.jubilant biosys.com).

[14] H. J. Chung, M. Kim, C. H. Park, J. Kim and J. H. Kim. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. Nucleic Acids Res, 32(Web Server issue):W460-4, 2004.

[15] H. J. Chung, C. H. Park, M. R. Han, S. Lee, J. H. Ohn, J. Kim, J. Kim and J. H. Kim. ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. Nucleic Acids Res, 33(Web Server issue):W621-6, 2005.

[16] Y. Xia, A. Campen, D. Rigsby, Y. Guo, X. Feng, E. W. Su, M. Palakal and S. Li. DGEM--a microarray gene expression database for primary human disease tissues. Mol Diagn Ther, 11(3):145-9, 2007.

[17] D. Cavalieri, C. Castagnini, S. Toti, K. Maciag, T. Kelder, L. Gambineri, S. Angioli and P. Dolara. Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. Bioinformatics, 23(19):2631-2, 2007.

[18] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor and B. R. Conklin. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet, 31(1):19-20, 2002.

[19] N. Salomonis, K. Hanspers, A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. Stuart, B. R. Conklin and A. R. Pico. GenMAPP 2: new features and resources for pathway analysis. BMC Bioinformatics, 8:217, 2007.

[20] J. Wu, X. Mao, T. Cai, J. Luo and L. Wei. KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res, 34(Web Server issue):W720-4, 2006.

[21] N. Goffard and G. Weiller. PathExpress: a web-based tool to identify relevant pathways in gene expression data. Nucleic Acids Res, 35(Web Server issue):W176-81, 2007.

[22] N. Goffard, T. Frickey and G. Weiller. PathExpress update: the enzyme neighbourhood method of associating gene-expression data with metabolic pathways. Nucleic Acids Res, 37(Web Server issue):W335-9, 2009.

[23] D. Pan, N. Sun, K. H. Cheung, Z. Guan, L. Ma, M. Holford, X. Deng and H. Zhao. PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. BMC Bioinformatics, 4:56, 2003.

[24] M. P. van Iersel, T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin and C. Evelo. Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics, 9:399, 2008.

[25] B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo and Z. Trajanoski. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Res, 33(Web Server issue):W633-7, 2005.

[26] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and R. Romero. A systems biology approach for pathway level analysis. Genome Res, 17(10):1537-45, 2007.

[27] R. Pandey, R. K. Guru and D. W. Mount. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. Bioinformatics, 20(13):2156-8, 2004.

[28] P. Grosu, J. P. Townsend, D. L. Hartl and D. Cavalieri. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. Genome Res, 12(7):1121-6, 2002.

[29] Z. Hu, J. Mellor, J. Wu and C. DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics, 5:17, 2004.

[30] Z. Hu, E. S. Snitkin and C. DeLisi. VisANT: an integrative framework for networks in systems biology. Brief Bioinform, 9(4):317-25, 2008.

[31] I. Tsui, R. Chari, T. Buys and W. Lam. Public Databases and Software for the Pathway Analysis of Cancer Genomes. Cancer Inform, 3:389-407, 2007.

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

[33] I. Vastrik, P. D'Eustachio, E. Schmidt, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney and L. Stein. Reactome: a knowledge base of biologic pathways and processes. Genome Biology, 8(3):R39, 2007.

[34] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi. KEGG for linking genomes to life and the environment. Nucleic Acids Res, 36(Database issue):D480-4, 2008.

[35] S. Krupa, K. Anthony, J. R. Buchoff, M. Day, T. Hannay and C. F. Schaefer. The NCI-Nature Pathway Interaction Database: A cell signaling resource. NCI-Nature Pathway Interaction Database, 2007. doi:10.1038/npre.2007.1311.1

[36] Biocarta, Charting Pathways of Life (http://www.biocarta.com).

[37] G. Jr. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol, 4(5):P3, 2003.

[38] F. Al-Shahrour, P. Minguez, J. Tárraga, I. Medina, E. Alloza, D. Montaner and J. Dopazo, FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Res, 35(Web Server issue):W91-6, 2007.

[39] J. M. Bland and D. G. Altman. Multiple significance tests: the Bonferroni method. BMJ, 310(6986):1073, 1995.

[40] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci, 102(43):15545-50, 2005.

[41] Prefuse, information visualization toolkit (http://prefuse.org).

[42] I. Cases, D. G. Pisano, E. Andres, A. Carro, J. M. Fernández, G. Gómez-López, J. M. Rodriguez, J. F. Vera, A. Valencia and A. M. Rojas. CARGO: a web portal to integrate customized biological information. Nucleic Acids Res, 35(Web Server issue):W16-20, 2007.