Published online 5 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W105–W109***
doi: 10.1093/nar/gkw356

PathwAX: a web server for network crosstalk based pathway annotation

Christoph Ogris[1,*], Thomas Helleday[2] and Erik L.L. Sonnhammer[1,*]

[1]Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden and [2]Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

Received February 10, 2016; Revised April 18, 2016; Accepted April 19, 2016

ABSTRACT

Pathway annotation of gene lists is often used to functionally analyse biomolecular data such as gene expression in order to establish which processes are activated in a given experiment. Databases such as KEGG or GO represent collections of how genes are known to be organized in pathways, and the challenge is to compare a given gene list with the known pathways such that all true relations are identified. Most tools apply statistical measures to the gene overlap between the gene list and pathway. It is however problematic to avoid false negatives and false positives when only using the gene overlap. The pathwAX web server (http://pathwAX.sbc.su.se/) applies a different approach which is based on network crosstalk. It uses the comprehensive network FunCoup to analyse network crosstalk between a query gene list and KEGG pathways. PathwAX runs the BinoX algorithm, which employs Monte-Carlo sampling of randomized networks and estimates a binomial distribution, for estimating the statistical significance of the crosstalk. This results in substantially higher accuracy than gene overlap methods. The system was optimized for speed and allows interactive web usage. We illustrate the usage and output of pathwAX.

INTRODUCTION

Functional genomics experiments are widely used to gain insights into biological processes. A typical experiment measures gene expression in a specific (perturbed) condition and a control from which a list of differentially expressed genes is calculated. This list does not normally give much direct insight as the differentially expressed genes may represent a mix of different biochemical functions and categories. However, if they are known to interact with each other in a pathway then this pathway is clearly affected.

A range of pathway databases exist (1), but the most general ones are Kyoto Encyclopedia of Genes and Genomes (KEGG) (2) and Gene Ontology (GO) (3). A large number of tools are available to assess whether a gene list is associated with a pathway or not (see (4) for a review). They typically apply a statistical measure such as the Fisher's exact test to assess the significance of the gene overlap between the gene list and pathway. However, because the pathway databases are far from complete, many true associations will be missed. Furthermore, the Fisher exact test generally overestimates the significance because it assumes that all genes are independent of each other, but this is not true as they interact with each other (5). The problem can be reduced with for instance the 'EASE score' (6) but still remains paramount. In summary, gene overlap methods produce high levels of false negatives and false positives, and there is a great need to improve this situation.

A solution has recently been proposed by network-based approaches such as NEA (7), EnrichNet (8), CrossTalkZ (9) and BinoX (Ogris *et al.*, submitted). These methods analyse enrichment of network links between gene sets rather than the gene overlap. If employing a dense comprehensive network of functional gene associations such as FunCoup (10,11) or STRING (12), the relation between gene list and pathway can be analysed using a lot more data than are provided by the gene overlap. Using crosstalk analysis one can also detect significant depletion of crosstalk relative to what is expected, which is never possible with gene overlap analysis.

However, network-based approaches face two main challenges: (1) which statistical model to assess significance, and (2) if employing iterative network rewiring to estimate a null model, how to avoid an excessive compute time? EnrichNet solves the latter by random walk with restart instead of randomizing the whole network. It however does not assess the statistical significance of the crosstalk enrichment. NEA and CrossTalkZ both use the normal distribution as model,

*To whom correspondence should be addressed. Tel: +46 70 5586395; Email: erik.sonnhammer@scilifelab.se
Correspondence may also be addressed to Christoph Ogris. Tel: +46 70 5586395; Email: christoph.ogris@scilifelab.se

but this is often unsuitable and leads to a high false positive rate for NEA and a high false negative rate for CrossTalkZ.

The binomial distribution, which is used by BinoX, is more suitable and gives much lower false positive and false negative rates (Ogris *et al.*, 2016, submitted). The BinoX algorithm employs Monte-Carlo sampling of randomized networks (while preserving topological properties) to estimate parameters of a binomial distribution which is used to calculate the statistical significance of an observed crosstalk. Two gene sets are considered to have significant enrichment if the number of network connections is significantly higher in the real network than expected from the random model. If the groups have fewer connections in the real network than expected from the random model, they have crosstalk depletion.

The earlier methods, CrossTalkZ and NEA, need to generate and analyse hundreds of randomized networks to calculate the statistical significance of a crosstalk for every query, and this is too time-consuming to be done interactively in a web site. To make it possible to obtain fast network crosstalk results, BinoX employs a very efficient statistical method based on pre-sampled randomized networks that are stored in a database, including information about up-to-date curated pathways.

To use BinoX, it is necessary to first obtain a large network and a set of pathways, and this can be a hurdle for some users. To make BinoX readily available to the public, we developed the pathwAX (pathway analysis with crosstalk) web server. It contains KEGG pathway information and genome-wide association networks of 11 well-studied model organisms. To minimize compute time, we have pre-randomized the networks, which gives run times for single gene sets of half a minute up to a few minutes. The pathwAX web site thus provides interactive online network crosstalk based pathway annotation that has a high chance of discovering affected pathways. We here illustrate this process with an example gene set containing 14 genes.

IMPLEMENTATION

pathwAX was designed to maximize performance and usability for network crosstalk based pathway annotation. The system relies on loading data dynamically allowing a multithreaded interplay between client and server modules. The server is running python 2.7 cgi scripts optimized for fetching and serving data, while the client side (the browser) is used for integrating data and estimating the statistical significance of crosstalk. The web service is based on javascript using the libraries jquery v2.1.4 and jstat for efficient data handling and calculation. For visualization, the libraries D3 v3.5.16 and Materialize v0.97.3 are used. The platform is optimized for the chrome browser; slightly more compute time should be expected using the browsers Firefox, Safari, Edge or Opera. Due to lack of support for some of the used javascript libraries, pathwAX is not compatible with Internet Explorer.

The BinoX workflow was adapted for pathwAX, where an annotation request is divided into four main stages, see Figure 1. During the first stage, pathwAX translates the query genes to internal IDs using the FunCoup web service, and a subnetwork including valid query genes and their

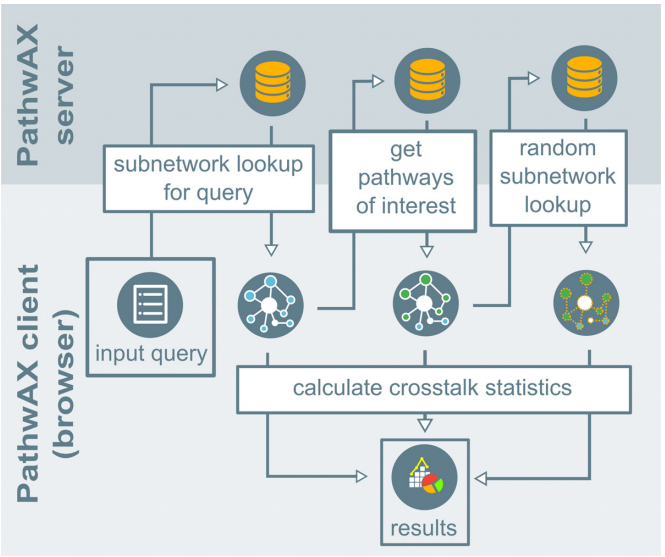

Figure 1. PathwAX workflow. After the user submits an input query, a subnetwork containing all query genes and their neighbours are requested from the server. A second request looks up all pathways sharing at least one gene with the subnetwork. In the final call the browser gets the parameters of the randomized connections between the pathways and the query gene. Once the browser has obtained the subnetwork, pathways of interest and the randomized connection parameters, it calculates the crosstalk statistics and displays these in the browser.

adjacent network genes is returned to the client. The second stage includes requesting relevant pathways, i.e. pathways having at least one connection to the query within the present subnetwork, as well as the total number of outgoing connections for the query gene set and each pathway. These are needed later for statistics. After combining the pathway information with the subnetwork, it is possible to count the number of network connections k between the query gene set and each pathway. In the third step, the client requests for each query-pathway pair the average number of connections k' in the randomized networks, which is used as an estimate of the expected connections within a randomized environment, $E(k')$. The final stage is initiated once all information is gathered. Using the BinoX algorithm the pathwAX client assumes the binomial distribution to employ alternative hypothesis testing to calculate the statistical significance of observing k. The binomial distribution for the alternative hypothesis depends on n', the maximum possible connections between gene set and pathway, and p', the probability of observing k'. Here p' can be approximated by $E(k')/n'$. Finally, pathwAX uses the Benjamini-Hochberg procedure to account for multiple testing and calculates corrected False Discovery Rate (FDR) values.

During each stage, pathwAX requests additional information of gene IDs, pathway names, pathway components, etc. The additional information is rendered together with the estimated FDR in the final summary to give the user an overview of the pathway annotation.

DATA

PathwAX incorporates 1930 pathways from the KEGG database (release 70.1) distributed among 11 model organ-

isms (see Table 1 for details). The FunCoup database provides a genome-wide functional association network for each species (link confidence score > 0.8). The networks are undirected, scale free and dense; they were constructed by integrating nine different evidence types: mRNA co-expression, phylogenetic profile similarity, protein–protein interaction, subcellular co-localization, co-microRNA regulation, domain interaction, protein co-expression, shared TF binding and genetic interaction profile similarity. Across all species' networks, an average gene has 90 network connections. The densest network is *Arabidopsis* with 154 neighbours on average for each gene whereas yeast is the sparsest with 45 links on average. Using the BinoX algorithm, 1000 randomized networks were sampled for each species. PathwAX supports a variety of different identifier types including Ensembl gene, protein as well as transcript IDs, NCBI gene IDs, RefSeq IDs and UniProt IDs.

USING PATHWAX

The input of pathwAX is a list of genes with identifier types as above, and the output is a table of significantly enriched or depleted KEGG pathways, together with graphical displays of pathway categories and the network connections between query genes and each pathway as a matrix. The gene list input size is limited to 400 as larger sizes compromises the rendering of the network connectivity matrix which shows graphically each query gene's connectivity to each pathway and provides hyperlinks to view these links in the FunCoup network. We want to ensure that this key functionality of the web site works.

To illustrate pathwAX, we chose an example from an experiment with a well-known phenotype: the gene list LOPEZ_MESOTHELIOMA_SURVIVAL_WORST_VS_BEST_UP (13) from MSigDB v3.0 collection (14). These 14 genes were expressed higher in the worst 25 survivors compared to the 25 best survivors, in a study of 99 pleural mesothelioma patients, and thus reflect processes that make this cancer more lethal. PathwAX identified 31 significant (FDR $\leq$ 0.05) pathways of which only 15 had any gene overlap (1 or 2 genes) with the query set, see Figure 2. The largest category, with 12 pathways, is Human Diseases, and four of these are cancer pathways. This particular form of cancer affects cells of the pleural mesothelium, the protective lining of the lungs and chest wall. It is therefore not surprising that many of the significant pathways are related to cell adhesion, including the most significant (Focal adhesion) and the third most significant (Leukocyte transendothelial migration). The latter pathway received an FDR of 8.9e-9, yet it would not be discoverable using traditional gene overlap methods as there is no overlap between it and the query gene list. This pathway contains 117 genes; 26 network links were observed between it and the query gene list, yet only 4.8 links are expected by chance. It is of clinical interest that four pathways related to heart disease have significant crosstalk with this gene list, which manifests poor survival.

DISCUSSION

In this paper we present pathwAX, a new web server for pathway annotation based on network crosstalk. The statistical model used for calculating significance is based on the binomial distribution. This increases sensitivity, accuracy and speed, and allows to combine network crosstalk pathway analysis with interactive web usage.

Compared to traditional gene overlap enrichment analysis, pathwAX offers a substantial improvement in sensitivity and specificity. For instance, applying gene overlap enrichment analysis using the EASE score (6), which is employed by the popular DAVID web site (15), to the example in Figure 2 is only possible for three of the 31 pathways found by pathwAX. The other 28 pathways cannot be tested because they have a gene overlap of less than two. Two is the minimum overlap to calculate the EASE score, which performs the Fisher exact test on the overlap minus 1. This is done in order to reduce the high false positive rate of the pure Fisher exact test.

A major problem with gene overlap enrichment analysis is that pathway annotation is incomplete. Only about a third of all human genes are annotated in KEGG, hence the chance of an overlap is small. Usually more than 90% of all pathway–gene set pairs overlap by less than two genes, and can therefore not be tested with the EASE score. In the example in Figure 2, the three pathways that can be tested with EASE had an overlap of exactly two genes. However, such a small overlap can often happen by chance, which would result in false positives.

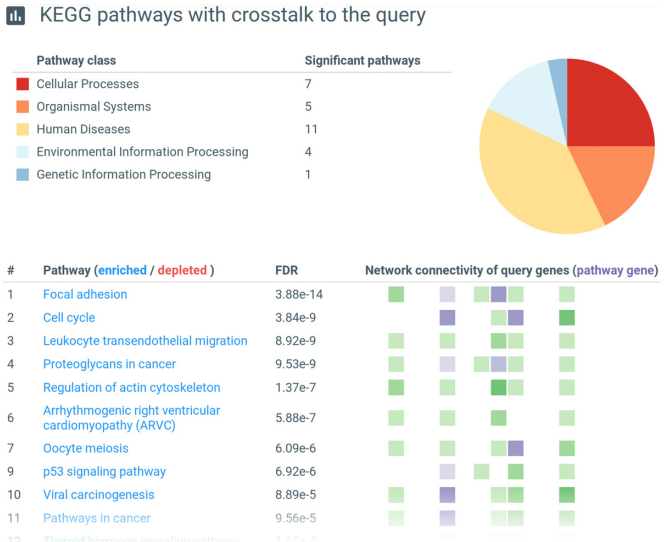

Figure 2. PathwAX results for the 14 genes in the human gene set LOPEZ_MESOTHELIOMA_SURVIVAL_WORST_VS_BEST_UP in MSigDB (PLXNA3, PSRC1, DLGAP4, HN1, CDC25C, FLNB, C20ORF20, CCND1, ACOT7, FLJ20674, TGFB1I1, LOX, DDAH1, CDC42EP3). The upper table and pie diagram summarize the results and visualize the distribution of pathway classes. The lower table lists all enriched (blue) and depleted (red) pathways for the query that are significant for the chosen cutoff (only top part shown). The pathways may be restricted to a class by clicking on one in the upper table. The results are sorted by increasing FDR. To the right is a matrix showing network connections between query genes and each pathway. Each gene is shown as a coloured box and mouseover shows its number of links to the pathway. Green boxes represent query genes linked to the pathway and purple boxes indicate genes which are part of the pathway. Darker shades indicate higher connectivity.

Table 1. Overview of networks and pathways available in pathwAX

Species	Network genes	Network connections	Pathways	Unique pathway genes
Homo sapiens	11 882	1 002 371	289	6 482
Mus musculus	12 903	1 495 536	286	7 299
Rattus norvegicus	12 025	1 668 050	271	6 458
Canis familiaris	9 292	667 556	244	4 712
Gallus gallus	6 211	299 485	135	3 210
Danio rerio	8 480	769 808	148	4 502
Ciona intestinalis	3 282	212 110	87	1 263
Drosophila melanogaster	5 762	385 691	124	2 395
Caenorhabditis elegans	6 014	686 340	124	2 014
Saccharomyces cerevisiae	3 991	179 499	101	1 784
Arabidopsis thaliana	9 306	1 433 523	121	4 239

Network genes are defined as protein coding genes having at least one connection within the network. The number of unique pathway genes relates to genes included in FunCoup.

PathwAX can analyse both enriched and depleted network crosstalk. What does a significantly depleted crosstalk imply? Technically it means that there are significantly fewer links than expected, and the implication is that there is statistical evidence that the gene set is not affected by a depleted pathway. One should thus not misinterpret it as an indication that the pathway is 'turned off'. In the example in Figure 2, the pathway Oxidative Phosphorylation was significantly depleted. This makes sense because cancer cells are mostly performing anaerobic energy production by glycolysis instead of oxidative phosphorylation, which is the aerobic energy production mostly used in healthy cells. This is called the Warburg effect (16).

Possible future extensions to pathwAX include using other pathway databases and networks. We chose the FunCoup network because of its comprehensiveness, which is paramount for crosstalk analysis. Although a variety of pathway databases exist, KEGG has the advantage of well-defined and relatively distinct pathways, which gives results that are easy to interpret, and it has good coverage for the species that we support. Methodologically one could consider using gene expression values more beyond just to extracting a list of differentially expressed genes. Methods exist that use such values to weight the relations to pathways (17,18). However, most of the information is captured by the list of significant differentially expressed genes and not much can be gained from including less informative data. Also, absolute gene expression levels are highly variable and it may be unwise to trust these too much. In practice, methods that use expression profiles are not as widely used as traditional gene overlap methods, possibly due to instability of the results and less interpretability. In a recent benchmark, most expression profile based methods did not show a clear advantage (17).

Another possibility would be to compare the pattern of crosstalk with the known wiring of genes within pathways to give higher weight to crosstalks with interacting genes. This may be a way to better rank the pathways, but we believe that the current level of biological knowledge is too fragmentary to dismiss a crosstalk based on poor consistency with the annotated wiring of the pathway.

SERVER INFORMATION

The web server is a virtual machine running Scientific Linux 6.7 with 2 GB RAM and 2 Intel Xeon E5-2630v2 2.60 GHz cores.

ACKNOWLEDGEMENT

We thank the Science for Life Laboratory for providing the infrastructure for the pathwAX web site.

FUNDING

Funding for open access charge: Swedish Research Council.
Conflict of interest statement. None declared.

REFERENCES

1. Ooi,H.S., Georg,S., Teng-Ting,L., Ying-Leong,C., Birgit,E. and Frank,E. (2010) Biomolecular pathway databases. *Methods Mol. Biol.*, **609**, 129–144.
2. Kanehisa,M., Minoru,K., Yoko,S., Masayuki,K., Miho,F. and Mao,T. (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
3. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
4. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
5. Gatti,D.M., Barry,W.T., Nobel,A.B., Rusyn,I. and Wright,F.A. (2010) Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, **11**, 574.
6. Hosack,D.A., Dennis,G. Jr, Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.
7. Alexeyenko,A., Lee,W., Pernemalm,M., Guegan,J., Dessen,P., Lazar,V., Lehtiö,J. and Pawitan,Y. (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, **13**, 226.
8. Glaab,E., Baudot,A., Krasnogor,N., Schneider,R. and Valencia,A. (2012) EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451–i457.
9. McCormack,T., Frings,O., Alexeyenko,A. and Sonnhammer,E.L.L. (2013) Statistical assessment of crosstalk enrichment between gene groups in biological networks. *PLoS One*, **8**, e54945.
10. Alexeyenko,A., Schmitt,T., Tjärnberg,A., Guala,D., Frings,O. and Sonnhammer,E.L.L. (2012) Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res.*, **40**, D821–D828.
11. Schmitt,T., Ogris,C. and Sonnhammer,E.L.L. (2014) FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res.*, **42**, D380–D388.

12. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
13. López-Ríos,F., Chuai,S., Flores,R., Shimizu,S., Ohno,T., Wakahara,K., Illei,P.B., Hussain,S., Krug,L., Zakowski,M.F. *et al.* (2006) Global gene expression profiling of pleural mesotheliomas: overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction. *Cancer Res.*, **66**, 2970–2979.
14. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
15. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
16. Koppenol,W.H., Bounds,P.L. and Dang,C.V. (2011) Otto Warburg's contributions to current concepts of cancer metabolism. *Nat. Rev. Cancer*, **11**, 325–337.
17. Dong,X., Xinran,D., Yun,H., Xiao,W. and Weidong,T. (2016) LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Sci. Rep.*, **6**, 18871.
18. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.