Published online 27 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W335–W339***
doi:10.1093/nar/gkp432

PathExpress update: the enzyme neighbourhood method of associating gene-expression data with metabolic pathways

Nicolas Goffard[1,2], Tancred Frickey[1] and Georg Weiller[1,*]

[1]ARC Centre of Excellence for Integrative Legume Research, Genomic Interactions Group, School of Biology, Australian National University, Canberra ACT 2601, Australia and [2]Institut Louis Malardé, Papeete, Tahiti, French Polynesia

Received February 1, 2009; Revised April 21, 2009; Accepted May 11, 2009

ABSTRACT

The post-genomic era presents us with the challenge of linking the vast amount of raw data obtained with transcriptomic and proteomic techniques to relevant biological pathways. We present an update of PathExpress, a web-based tool to interpret gene-expression data and explore the metabolic network without being restricted to predefined pathways. We define the Enzyme Neighbourhood (EN) as a sub-network of linked enzymes with a limited path length to identify the most relevant sub-networks affected in gene-expression experiments. PathExpress is freely available at: http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/.

INTRODUCTION

With the development of transcriptomic and proteomic techniques, post-genomic data represents a new challenge for researchers attempting to interpret the vast amount of raw data in a biological context (1). The analysis of microarray data is usually performed in two steps: the identification of genes that are differentially expressed under two or more conditions, using different statistical methods (2), and a comparison of selected genes with a background to find overlaps between the observed changes in expression and biologically relevant partitionings of the measured genes. Many ontological tools are now available that support the functional interpretation of gene-expression data via the identification of significantly enriched Gene Ontology (GO) categories (3) within groupings of genes of interest (4).

Additionally, with the availability of pathway databases such as KEGG (5,6) and MetaCyc (7), numerous tools have been proposed that analyse microarray data and visually present associated metabolic or regulatory pathway information (8–16). However, the predefined metabolic pathways used in these methods represent an essentially arbitrary segmentation of the metabolism. In contrast, other methods integrate, a priori, the knowledge of gene networks in the analysis of gene-expression data. Ideker and co-workers presented a procedure for screening a molecular interaction network combined with a statistical measure to identify sub-networks that show significant changes in expression (17). This approach has been included in Cytoscape to identify functional modules, i.e. highly connected network regions with similar responses across multiple experimental conditions (18). Hanisch and co-workers proposed a co-clustering method based on a distance function that combines information from expression data and biological networks (19). A Potts spin algorithm was developed to cluster gene-expression data by using the nearest neighbour reactions of biochemical networks (20). Rapaport and co-workers extracted gene-expression patterns of neighbouring genes in the network, involving the attenuation of high-frequency signals with respect to the graph (21). Another approach identifies the smallest functional units based on the network topology using the Petri net theory (22). It has been shown by Schwartz and co-workers that elementary modes represent true functional units of metabolism and can be used to reveal transcriptional activity (23). However, the combinatorial explosion of computing elementary modes in large networks limits the practical use of these methods.

We previously presented a web-based tool called PathExpress (10) that allowed us to interpret gene-expression results from microarrays in the context of biological pathways. PathExpress has been developed to identify the most relevant pathways or sub-pathways associated with a subset of genes of interest (e.g. a set of differentially expressed genes). It is based on a directed graph modelling enzymatic reactions derived from the publicly available KEGG LIGAND database (24,25).

*To whom correspondence should be addressed. Tel: +61 2 6125 5916; Fax: +61 2 6125 7879; Email: Georg.Weiller@anu.edu.au

In the present article, we describe a new development in PathExpress—the enzyme neighbourhood (EN) method. We define the EN as a sub-network of linked enzymes with a limited path length. The EN method enables us to explore the metabolic network and identify the most relevant sub-networks affected in gene-expression experiments without being restricted to predefined pathways. While the interaction with the web server is essentially unchanged, PathExpress now incorporates the EN method and supports 28 Affymetrix 3′ Gene-expression Analysis Arrays, representing 32 distinct organisms, and is easy to extend further. In a case study, the EN method was tested with gene-expression data of the model legume *Medicago truncatula* by comparing the transcriptomes of meristematic and non-meristematic root cells (26).

METHODS

Data representation

PathExpress is based on a directed graph modelling enzymatic reactions as used in the Petri net representation of biological networks (27). Two types of nodes are used to represent compounds and reactions. Specific reactions can encompass one or more enzymes. Directed edges, connecting these nodes, correspond to the consumption or the production of compounds by the reaction. We first built the global metabolic network consisting of 2276 enzymes and 3810 compounds involved in 3663 reactions as specified in the KEGG LIGAND database (24,25). In order to avoid annotation errors due to the misinterpretation of partial Enzyme Commission (EC) numbers (28), we only utilized enzymes defined by a full EC term. This database has the advantage of providing a manually curated representation of enzymatic reactions involved in metabolic pathways where most secondary metabolites (very common and highly connected compounds such as water, oxygen, major coenzymes and prosthetic groups) have been removed, thus avoiding invalid metabolic connections and unspecified pathways.

Many of the current methods for the functional interpretation of gene-expression data are constrained by their need to link expressed genes with predefined metabolic pathways and are therefore often hampered when the species to be analysed is not represented in the pathway database. To overcome this limitation, probe sets of the genome arrays supported in PathExpress are linked to the metabolic network using NetAffx annotations (29) or similarities with protein sequences of known EC numbers retrieved from the UniProt database (30). A complete metabolic graph representing all assignments is produced for each organism. This strategy can be applied to any set of sequences and makes it easy to extend PathExpress for use with novel species. In addition, EC numbers can be directly uploaded and compared to the reference network, which allows the analysis of custom data.

ENZYME NEIGHBOURHOOD

In the global network, two reactions are regarded as neighbours if a metabolite exists that is the product of one reaction and the substrate of the other. We define the EN of depth d for an enzyme e, as the set of enzymes that can be reached in the graph from e by traversing a maximum of d compounds, regardless of the direction of the edges (Figure 1). The EN of depth 1 for a given enzyme thus corresponds to the set of enzymes directly connected via a compound (e.g. immediate neighbours). The EN of depth 2 includes the enzymes involved in the EN of depth 1 plus the enzymes linked to these. As different paths can connect two enzymes, the shortest distance between two enzymes is used to define the EN. These ENs correspond to different sub-networks of the global metabolic network. By comparing a specific list of genes to the ENs it is possible to identify those ENs that are significantly over-represented in the gene list.

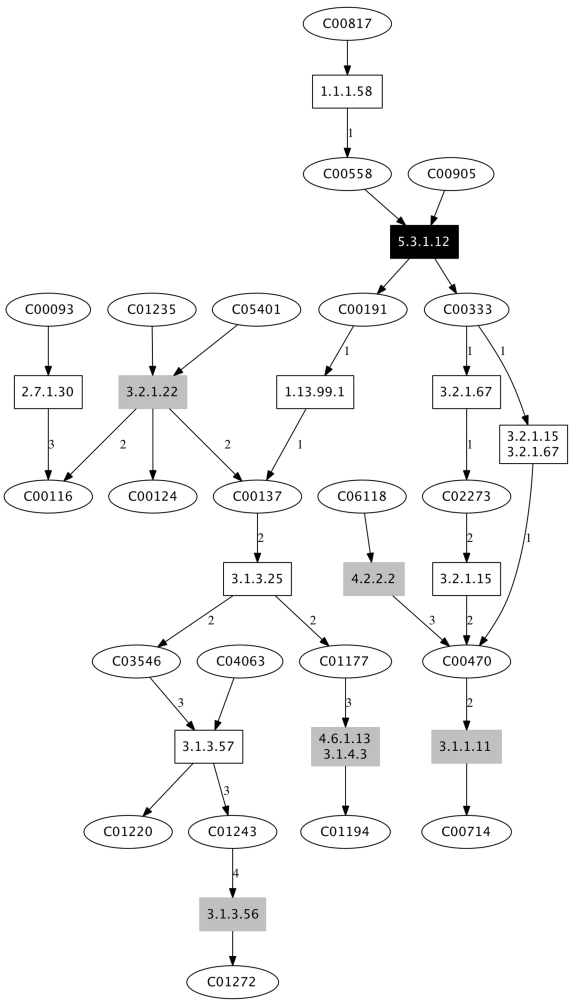

Figure 1. EN of depth 4, identified from a list of differentially expressed genes in *Medicago truncatula*. Compounds (labelled with their KEGG identifier and represented as ellipses) and reactions (labelled with the EC number of the enzymes that mediate it and represented as boxes) are the nodes of the directed graph. The enzyme coloured in black was used to seed this EN (entry point). Greyed reactions show that at least one enzyme thought to be capable of catalyzing the corresponding reaction was present in the submitted list of genes. The label of edges indicates the level of EN depth, i.e. the minimal number of compounds traversed in the global network from the seed enzyme to this point, regardless of the direction of the edges.

Table 1. Average size of the EN according to the depth parameter

Depth	Average no. of neighbours
1	11.7
2	14.5
3	21.9
4	34.0
5	51.0
6	74.2
7	105.5
8	145.1
9	193.8
10	253.5
20	995.0
30	1397.7
40	1622.1
50	1767.4
100	2106.8

To identify the most relevant sub-network associated with a list of submitted enzymes, the EN of each seed (submitted EC number), for a given depth, is determined in the global network and the EC numbers contained in the resulting EN are compared to the submitted list. For each test, a *P*-value, representing the probability that the intersection of the list of enzymes belonging to the given EN occurs per chance in the population of enzymes involved in the entire network, is calculated using the hypergeometric distribution (31). Because multiple tests are performed, it is necessary to correct these *P*-values with adjustment methods such as the conservative Bonferroni correction (32) or the False Discovery Rate approach (33).

The size of the EN depends on its depth *d*, which has to be specified as a parameter in the current implementation. To optimize this parameter with the size of the submitted list of genes, we have computed the average number of enzymes involved in each possible EN for a range of depths (Table 1). Based on these results, it is possible to adjust the depth parameter to compare groups of enzymes with sub-networks of similar size. For example, to compare a group of 10 enzymes, we recommend a depth parameter of 1 (i.e. direct neighbours), corresponding to an average size of 11.7 enzymes.

THE PATHEXPRESS WEB SERVER

As input data, PathExpress receives a list of identifiers (Affymetrix probe set identifiers and/or GenBank accession numbers). Other parameters can be specified: the type of comparison (pathway, sub-pathway or EN), the *P*-value significance threshold and the adjustment method used to correct for multiple testing.

The PathExpress output contains the list of sub-networks (metabolic pathways, sub-pathways or ENs) that are associated with the enzymes in the submitted list of identifiers. The ones with significant association are highlighted. Each of these networks can be displayed, both via an automatically generated graphical representation and as an enumeration of enzymatic reactions.

APPLICATION EXAMPLE

As an example, we used PathExpress to analyse microarray data obtained from the model legume *Medicago truncatula*, comparing the gene expression of meristematic and non-meristematic root tissues (26). The data have been deposited in NCBI's Gene Expression Omnibus (34) and are accessible through GEO series accession number GSE8115. Following normalization, differentially expressed probe sets were identified by evaluating the $\log_2$ ratio between the two conditions. All probe sets that differed by more than a 2-fold difference were considered to be differentially expressed. Of the 390 transcripts over-expressed in the non-meristematic tissue, 94 could be assigned to 50 distinct enzymatic functions, as defined by their EC number in the Affymetrix Medicago Genome Array. To contrast the whole pathway approach with the EN method, we used the 'Entire Pathway' option of PathExpress to identify over-representation of metabolic pathways in the non-meristematic root. Most significantly (*P*-value: 1.09*e*–03), the carbon fixation pathway is defined by 22 enzymes of which six are differentially expressed in the tissue. We also identified the most relevant sub-networks corresponding to the same group of over-expressed transcripts, using the EN option with a depth of 4. The resulting sub-networks were ranked by increasing *P*-values. The most significant EN (*P*-value: 4.06*e*–04) is given in Figure 1 and was seeded by the glucuronate isomerase (EC 5.3.1.12, black). Of the 13 enzymes present in the depicted sub-network, seven are involved in the pentose and glucuronate interconversion pathway as described in the KEGG database. The remaining six enzymes connected to this sub-network are part of different pathways involved in carbohydrate metabolism (galactose, inositol phosphate, ascorbate and aldarate) and would not have been considered by an approach restricted to the predefined metabolic pathways.

DISCUSSION

Our web-based tool for the interpretation of genomics data, first described in 2007 (10), has been extended to implement the concept of ENs. The EN of a given enzyme is defined as a connected sub-network within the global metabolic network, built from the KEGG database. The identification of statistically significantly over-represented ENs is based on the same statistical approach used for the identification of gene enrichment in GO terms or metabolic pathways. However, the clustering method differs, as it includes knowledge about the network of gene products without being restricted to predefined pathways.

Recently, another tool called KEGG spider, presenting a similar approach of interpretation of genomics data in the context of the global gene metabolic network, has been published (35). Although both methods identify statistically significant sub-networks in a submitted list of

genes, there are some fundamental differences. KEGG spider infers the network that minimizes the distance between each connected gene pair according to pair-wise distances between genes. It estimates the significance of the inferred network by a Monte Carlo procedure. On the other hand, PathExpress performs an enrichment analysis by comparing the EN of a given depth with the submitted genes, using the hypergeometric distribution and an adjustment method. While KEGG spider limits sub-networks by allowing a maximum of three consecutive missing enzymes, PathExpress can consider all sub-networks up to a depth of 10, corresponding to approximately 250 enzymes. KEGG spider uses the KEGG orthology database to map the genes to the metabolic network and is available only for nine reference organisms, whereas PathExpress uses pre-computed assignments of sequences to EC numbers, and can easily be extended from the currently supported 32 organisms to any organism or set of sequences (e.g. custom DNA microarray, proteome array), enabling the analysis of a wider range of gene-expression experiments. For example, it has recently been used to compare the proteomic data derived from seeds of plants within and beyond the legume family (36).

Since its initial development, PathExpress has been extended to explore the Enzyme Neighbourhood for the identification of relevant sub-networks affected in gene-expression experiments. Many genome arrays have been added, making PathExpress a useful resource for the integration of transcriptomic and proteomic and enzymatic or metabolic reaction datasets.

FUNDING

Australian Research Council Centre of Excellence Grant. Funding for open access charge: Australian Research Council Centre of Excellence Grant.

Conflict of interest statement. None declared.

REFERENCES

1. Breitling,R. (2006) Biological microarray interpretation: the rules of engagement. *Biochim. Biophys. Acta*, **1759**, 319–327.
2. Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.
3. Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
4. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
5. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
6. Okuda,S., Yamada,T., Hamajima,M., Itoh,M., Katayama,T., Bork,P., Goto,S. and Kanehisa,M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, **36**, W423–W426.
7. Caspi,R., Foerster,H., Fulcher,C.A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J., Rhee,S.Y. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.
8. Baitaluk,M., Sedova,M., Ray,A. and Gupta,A. (2006) BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res.*, **34**, W466–W471.
9. Chung,H.J., Park,C.H., Han,M.R., Lee,S., Ohn,J.H., Kim,J., Kim,J. and Kim,J.H. (2005) 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**, W621–W626.
10. Goffard,N. and Weiller,G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res.*, **35**, W176–W181.
11. Mlecnik,B., Scheideler,M., Hackl,H., Hartler,J., Sanchez-Cabo,F. and Trajanoski,Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, **33**, W633–W637.
12. Pan,D., Sun,N., Cheung,K.H., Guan,Z., Ma,L., Holford,M., Deng,X. and Zhao,H. (2003) PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for *Arabidopsis*. *BMC Bioinformatics*, **4**, 56.
13. Pandey,R., Guru,R. and Mount,D. (2004) Pathway miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, **20**, 2156–2158.
14. Salomonis,N., Hanspers,K., Zambon,A.C., Vranizan,K., Lawlor,S.C., Dahlquist,K.D., Doniger,S.W., Stuart,J., Conklin,B.R. and Pico,A.R. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.
15. Thimm,O., Bläsing,O., Gibon,Y., Nagel,A., Meyer,S., Krüger,P., Selbig,J., Müller,L.A., Rhee,S.Y. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.
16. Wu,J., Mao,X., Cai,T., Luo,J. and Wei,L. (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34**, W720–W724.
17. Ideker,T., Ozier,O., Schwikowski,B. and Siegel,A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18(Suppl. 1)**, S233–S240.
18. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
19. Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18(Suppl. 1)**, S145–S154.
20. König,R. and Eils,R. (2004) Gene expression analysis on biochemical networks using the Potts spin model. *Bioinformatics*, **20**, 1500–1505.
21. Rapaport,F., Zinovyev,A., Dutreix,M., Barillot,E. and Vert,J.P. (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.
22. Schuster,S., Pfeiffer,T., Moldenhauer,F., Koch,I. and Dandekar,T. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics*, **18**, 351–361.
23. Schwartz,J.M., Gaugain,C., Nacher,J.C., de Daruvar,A. and Kanehisa,M. (2007) Observing metabolic functions at the genome scale. *Genome Biol.*, **8**, R123.
24. Goto,S., Nishioka,T. and Kanehisa,M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.
25. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.
26. Holmes,P., Goffard,N., Weiller,G.F., Rolfe,B.G. and Imin,N. (2008) Transcriptional profiling of *Medicago truncatula* meristematic root cells. *BMC Plant Biol.*, **8**, 21.
27. Sackmann,A., Heiner,M. and Koch,I. (2006) Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, **7**, 482.
28. Green,M.L. and Karp,P.D. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.*, **33**, 4035–4039.
29. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003)

NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
30. The UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
31. Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W. and Lockhart,D.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.
32. Bonferroni,C. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.
33. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, **57**, 289–300.
34. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
35. Antonov,A.V., Dietmann,S. and Mewes,H.W. (2008) KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol.*, **9**, R179.
36. Dam,S., Laursen,B.S., Ornfelt,J.H., Jochimsen,B., Staerfeldt,H.H., Friis,C., Nielsen,K., Goffard,N., Besenbacher,S., Krusell,L. *et al.* (2009) The proteome of seed development in the model legume *Lotus japonicus*. *Plant Physiol.*, **149**, 1325–1340.