Exploring the Enzyme Neighbourhood to interpret gene expression data

Nicolas Goffard, Tancred Frickey, Nijat Imin, Georg Weiller

ARC Centre of Excellence for Integrative Legume Research
Bioinformatics Laboratory, Genomic Interactions Group
Research School of Biological Sciences, Australian National University
GPO Box 475, Canberra, ACT 2601, Australia
ngoffard@gmail.com
tancred.frickey@anu.edu.au
nijat.imin@anu.edu.au
georg.weiller@anu.edu.au

Abstract: Post-genomic data analysis represents a new challenge to link and interpret the vast amount of raw data obtained with transcriptomic or proteomic techniques in the context of metabolic pathways. We propose a new strategy with the help of a metabolic network graph to extend PathExpress, a web-based tool to interpret gene expression data, without being restricted to predefined pathways. We defined the Enzyme Neighbourhood as groups of linked enzymes, corresponding to a sub-network, to explore the metabolic network in order to identify the most relevant sub-networks affected in gene expression experiments.

1 Introduction

With the development of transcriptomic and proteomic techniques, post-genomic data analysis represents a new challenge for researchers to link the vast amount of raw data to a biological context [Br06]. The interpretation of microarray data is usually performed in two steps. The first step is the identification of genes that are differentially expressed under two or more conditions, using different statistical methods [CC03]. In a second step, the selected genes are compared with a background in order to find enrichment in any functional term. Many ontological tools are now available that support the functional interpretation of gene expression data, through the identification of significantly enriched Gene Ontology categories [As00] among a class of genes of interest [KD05].

Additionally, with the availability of pathway databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [KG00] or MetaCyc [Ca06], numerous tools have been proposed to visualize and analyse microarray data in the context of known biological networks by including metabolic or regulatory pathway information [Pa03], [PGM04], [Th04], [Ch05], [Ml05], [Ba06], [Wu06], [GW07], [Sa07]. However, the predefined metabolic pathways used in these methods represent an arbitrary segmentation of 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 [Id02]. This approach has been included in Cytoscape to identify functional modules, i.e. highly connected network regions with similar responses across multiple experimental conditions [Cl07]. Hanisch and co-workers proposed a co-clustering method based on a distance function that combines information from expression data and biological networks [Ha02]. A Potts spin algorithm was developed to cluster gene expression data by using the nearest neighbour relations of biochemical networks [KE04]. 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 [Ra07]. Another approach consists of the development of techniques for the decomposition of biochemical networks into the smallest functional units based on the network topology using the Petri net theory [Sc02], [SHK06]. 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 [Sc07]. However, these methods are limited by the combinatorial explosion of computing elementary modes in large networks.

We recently presented a web-based tool called PathExpress [GW07] to interpret gene expression results from microarray experiments in the context of biological pathways, available at http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/. PathExpress has been developed to identify the most relevant pathways or sub-pathways associated with a subset of genes, e.g., differentially expressed. It is based on a directed graph to model enzymatic reactions, derived from the publicly available KEGG Ligand database of chemical compounds and reactions in biological pathways [GNT98], [Go02]. Two types of nodes are used to represent compounds and reactions that can be mediated by one or more enzymes. To take into account how reactions are linked in pathway, sub-pathways are defined as a chain of reactions linked to each other by a common compound (substrate or product). Thus, PathExpress compares a submitted list of genes to the genes involved in annotated pathways or sub-pathways and identifies the significantly over-represented set of enzymatic reactions in the query using a hypergeometric distribution [Ch01]. This statistical test has been employed by many ontological tools to detect significant enrichments of functional categories within a class of genes of interest [Ri07].

This article presents developments in PathExpress that explore the metabolic network for the interpretation of gene expression data. We created a graph representing the complete metabolic network, which allows us to examine the neighbourhood of a given enzyme by following the chain of connected reactions linked by a common compound. The Enzyme Neighbourhood (EN) represents a group of linked enzymes corresponding to a sub-network. The EN can then be compared to a submitted list of genes with the aim to find ENs in which the submitted genes are significantly over-represented. In a case study, our method was tested with gene expression data of the model legume *Medicago truncatula* to compare the transcriptomes of meristematic and non-meristematic root cells [Ho08].

2 Methods

This approach is based on a directed graph modelling enzymatic reactions as used in the Petri net representation of biological networks [SHK06]. Two types of nodes are used to represent compounds and reactions with reactions represented by 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 2,198 enzymes and 2,796 compounds involved in 3,706 reactions as specified in the KEGG LIGAND database [GNT98], [Go02]. 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.

In this network, two reactions are neighbours if a metabolite exists that is the product of one reaction and the substrate for the other. Then, we define the Enzyme Neighbourhood (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 (Fig 2.1). The EN of depth 1 for a given enzyme thus corresponds to the set of enzymes directly connected via a compound. The EN of depth 2 includes the enzymes involved in the EN of depth 1 plus the enzymes linked to them. As different paths can connect two enzymes, the shortest distance is considered to define the EN. These ENs correspond to different sub-networks of the global metabolic network.

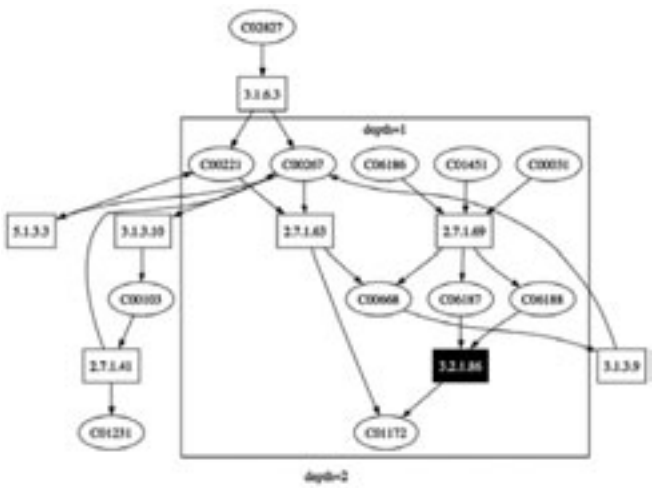

Figure 2.1: Example of an Enzyme Neighbourhood (EN). Compounds (labelled with their KEGG identifier and represented as ellipses) and reactions (labelled with the EC number of the enzymes that mediates it and represented as boxes) are the nodes of the directed graph. The EN of depth 1for the enzyme ‘EC 3.2.1.86’ contains the enzymes ‘EC 3.2.1.86’, ‘EC 2.7.1.69’ and ‘EC 2.7.1.63’, whereas the EN of depth 2 contains in addition to those includes in the EN of depth 1, the enzymes ‘EC 3.1.6.3’, ‘EC 5.1.3.3’, ‘EC 3.1.3.9’, ‘EC 2.7.1.41’ and ‘EC 3.1.3.10’.

To identify the most relevant sub-network associated with a list of submitted enzymes, the EN of each seed (EC number) is determined in the global network, for a given depth, and its composite EC numbers are compared to the submitted list. For each test, a *p*-value representing the probability that the intersection *k* of the list of enzymes of size *n* belonging to the given EN, of size *D*, occurs per chance in the population of *N* enzymes involved in the entire network, is calculated using the hypergeometric distribution [Ch01] as described below.

$$p(k,N,D,n) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

Because multiple hypothesis tests are performed, it is necessary to correct these *p*-values with adjustment methods such as the conservative Bonferroni correction [Bo06], in which the *p*-values are multiplied by the number of comparisons, or the less stringent False Discovery Rate (FDR) approach [BH95], which determines the expected proportion of false positive results among all rejected hypotheses.

The size *D* of the EN depends on its depth *d*, which has to be specified as a parameter in the current implementation. It is typically necessary to examine several ENs with different depths. To optimize this parameter with the size of the submitted list of enzymes, we have computed the average number of enzymes involved in each possible EN for a range of depths (Table 2.1). Using 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, a depth parameter of 1 (i.e. direct neighbours), corresponding to an average size of 11.7 enzymes in the network, is recommended.

Table 2.1: Average size of the Enzyme Neighbourhood 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

3 Application to gene expression data

We extended the web-based tool PathExpress with this method of exploring the Enzyme Neighbourhood in order to identify the most relevant sub-networks associated with a list of genes (e.g. differentially expressed genes).

3.1 Linking expressed enzymes with metabolic networks

One of the main constraints in methods for the functional interpretation of gene expression data corresponds to the linkage of such data to the metabolic network, as the number of available organisms in pathway databases is limited. To overcome this, we use similarities between probe set sequences of supported genome arrays and protein sequences of known EC numbers, retrieved from the Swiss-Prot database [Ba05], in order to link probe sets to the metabolic network (Table 3.1). Blastx [Al90] is used to find the best match (E-value $\leq 10^{-8}$) for the sequences representing each probe set sequence (i.e. sequences derived from the most 5' to the most 3' probe in the public Unigene cluster) of the genome arrays analyzed. If these entries have been annotated as enzymes, the probe set is assigned to the corresponding EC number, extracted from its definition line. This strategy can be applied to any set of sequences. A complete metabolic graph representing all assignments is produced and all qualifying sub-networks are compared with the data of a submitted genome array. High scoring Enzyme Neighbourhoods are then presented.

Note that probe sets that cannot be assigned to EC numbers are excluded from further analyses, and although this limits the number of usable probe sets, it also eliminates non-enzymatic gene functions that are present in many unrelated metabolic pathways. As the comparisons are based on enzyme composition rather than single probe set assignments, biases that arise from a multiplicity of genes coding for the same enzyme are largely overcome and the functional activities become apparent.

Table 3.1: Available Affymetrix genome arrays and assignment statistics

Affymetrix Genome Array (Organism)	% Sequences assigned	No. of ECs	No. of reactions
ATH1 Genome Array (*A. thaliana*)	22.7	823	1,177
E. coli Genome 2.0 Array (*E. coli*)	22	803	1,217
Drosophila Genome 2.0 Array (*D. melanogaster*)	16.4	724	1,011
Yeast Genome 2.0 Array (*S. cerevisiae*)	25.3	601	918
Yeast Genome 2.0 Array (*S. pombe*)	26.5	566	839
Medicago Genome Array (*M. truncatula*)	17.6	953	1,412
Soybean Genome Array (*G. max*)	17.2	803	1,217
Rice Genome Array (*O. sativa*)	17.6	923	1,363

3.2 Microarray data analysis

Our method was applied to interpret a microarray experiment in the model legume *Medicago truncatula*, comparing the gene expression of meristematic and non-meristematic root tissues [Ho08]. The data have been deposited in NCBI's Gene Expression Omnibus [EDL02] and are accessible through GEO series accession number GSE8115 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8115). Following normalisation, differentially expressed probe sets were identified by evaluating the log2 ratio between the two conditions associated to a standard *t*-test [Ca00]. All probe sets that differed by more than a two-fold difference with a *t*-test $p \leq 0.05$ were considered to be differentially expressed. Of the 363 transcripts over-expressed in the non-meristem, 119 could be assigned to 62 different enzymatic functions, defined by their EC number and found in the Affymetrix Medicago Genome Array. In order to identify the most relevant sub-networks involved in this group, we compare it, using PathExpress, to all ENs with a depth of 6, using the hypergeometric distribution. The resulting sub-networks were ranked by increasing *p*-values, representing the probability that the intersection of the enzymes differentially expressed in the non-meristem with the given EN occurs by chance.

The most significant EN (*p*-value = 1.4e-4), using the flavonone 3-dioxygenase (EC 1.14.11.9) as seed (black), is given in Figure 3.1. Of the 20 enzymatic reactions present in the depicted sub-network, 9 occur in the submitted list of differentially expressed enzymes (grey and black). Only 12 of the 20 reactions in this EN are part of the classical flavonoid biosynthesis pathway as described in the KEGG database, which is consistent with the role for the flavonoids and their derivatives in the non-meristematic root [Im07]. The remaining 8 reactions connected to this sub-network are part of different pathways (such as propanoate metabolism or liminene and pinene degradation) and would not have been considered by an approach restricted to predefined metabolic pathways.

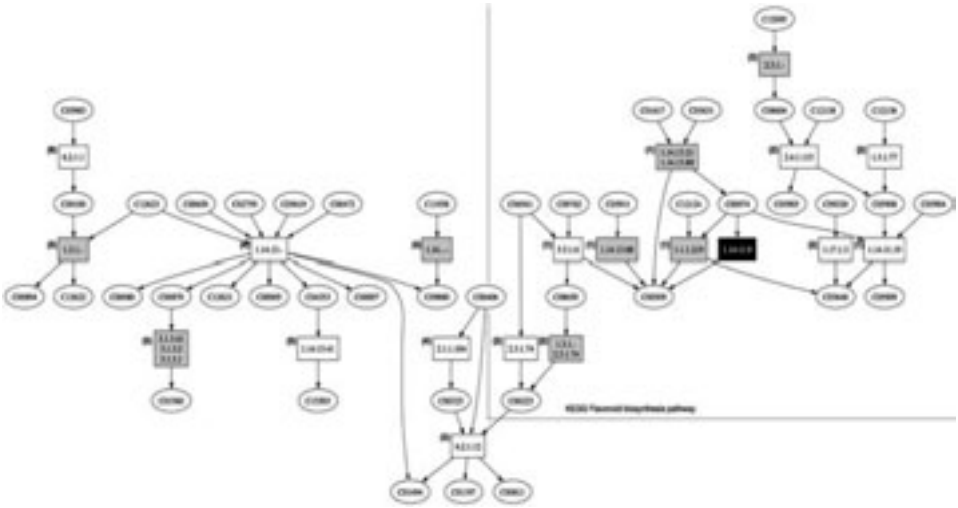

Figure 2.1: Enzyme Neighbourhood of depth 6, identified from a list of differentially expressed genes in *Medicago truncatula*. For each reaction represented, the EN depth is indicated (number in brackets). Genes encoding enzymes for all these reactions have been identified in the Affymetrix Medicago Genome Array. The reaction coloured in black corresponds to the enzyme (EC 1.14.11.9) used to establish this EN. Greyed reactions show that at least one of the corresponding enzymes belongs to the submitted group of enzymes. The set of reactions inside the frame represent part of the classical flavonoid biosynthesis pathway as described in KEGG database.

4 Conclusion

The interpretation of microarray experiments represents a main challenge to characterize biological processes. This paper presents a method to interpret results of gene expression data in the context of metabolic pathways. Our web-based tool PathExpress, in which metabolic pathways are modelled as directed graphs of enzymatic reactions, has been extended to identify Enzyme Neighbourhoods (EN) with statistically significant differential expressions. The EN of a given enzyme is defined as a connected sub-network within the global metabolic network, built from the KEGG database. This method is based on the same statistical approach as 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. Based on a pre-computed assignment of sequences to EC numbers this approach can be applied to any organism or set of sequences (e.g. custom DNA microarray, proteome array) and hence provides a useful resource for the integration of transcriptomic and proteomic data sets.

References

[Al90] Altschul, S.F. et. al.: Basic local alignment search tool. *J. Mol. Biol.*, 1990; vol. 215(3), pp. 403-410.

[As00] Ashburner, M. et. al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 2000; vol. 25, pp. 25-29.

[Ba05] Bairoch, A. et. al.: The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 2005; vol. 33(Database issue), pp. D154-159.

[Ba06] Baitaluk, M. et. al.: BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res*., 2006; vol. 34(Web Server issue), pp. W466-471.

[BH95] Benjamini, Y.; Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.,* 1995; vol. 57(1), pp. 289 - 300.

[Bo36] Bonferroni, C.: Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze,* 1936; vol. 8, pp. 3-62.

[Br06] Breitling, R.: Biological microarray interpretation: the rules of engagement. *Biochim. Biophys. Acta.*, 2006; vol. 1759(7), pp. 319-327.

[Ca00] Callow, M.J. et. al.: Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, 2000; vol. 10, pp. 2022-2029.

[Ca06] Caspi, R. et. al.: MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*., 2006; vol. 34(Database issue), pp. D511-516.

[CC03] Cui, X.; Churchill, G.A.: Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, 2003; vol. 4(4), pp. 210.

[Ch01] Cho, R.J. et. al.: Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, 2001; vol. 27(1), pp. 48-54.

[Ch05] Chung, H.J. et. al.: ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*., 2005; vol. 33(Web Server issue), pp. W621-626.

[Cl07] Cline, M.S. et. al.: Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, 2007; vol. 2(10), pp. 2366-2382.

[EDL02] Edgar, R.; Domrachev, M.; Lash, AE.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res., 2002; vol*. 30(1), pp. 207-210.

[Go02] Goto, S. et. al.: LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*., 2002; vol. 30(1), pp.402-404.

[GNT98] Goto, S.; Nishioka, T.; Kanehisa, M.: LIGAND: chemical database for enzyme reactions. *Bioinformatics*, 1998; vol. 14(7), pp. 591-599.

[GW07] Goffard, N.; Weiller, G.: PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res*., 2007; vol. 35(Web Server issue), pp. W176-181.

[Ha02] Hanisch, D. et. al.: Co-clustering of biological networks and gene expression data. *Bioinformatics*, 2002; vol. 18 Suppl 1, pp. S145-154.

[Ho08] Holmes, P. et. al.: Transcriptional profiling of Medicago truncatula meristematic root cells. *BMC Plant Biol*., 2008; vol. 8(1), pp. 21.

[Id02] Ideker, T. et. al.: Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 2002; vol. 8 Suppl 1, pp. S233-240.

[Im07] Imin, N. et. al.: Factors involved in root formation in Medicago truncatula. *J. Exp. Bot.*, 2007; vol. 58(3), pp. 439-451.

[KD05] Khatri, P.; Drăghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 2005; vol. 21(18), pp. 3587-3595.

[KE04] König, R.; Eils, R.: Gene expression analysis on biochemical networks using the Potts spin model. *Bioinformatics*, 2004; vol. 20(10), pp. 1500-1505.

[KG00] Kanehisa, M.; Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 2000; vol. 28, pp. 27-30.

[Ml05] Mlecnik, B. et. al.: PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, 2005; vol. 33(Web Server issue), pp. W633-637.

[Pa03] Pan, D. et. al.: PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. *BMC Bioinformatics*, 2003; vol. 4, pp. 56.

[PGM04] Pandey, R.; Guru, R.; Mount, D.: Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, 2004; vol. 20, pp. 2156-2158.

[Ra07] Rapaport, F. et. al.: Classification of microarray data using gene networks. *BMC Bioinformatics*, 2007; vol. 8, pp. 35.

[Ri07] Rivals, I. et. al.: Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 2007; vol. 23(4), pp. 401-407.

[Sa07] Salomonis, N. et. al.: GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 2007; vol. 8, pp. 217.

[Sc02] Schuster, S. et. al.: Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, 2002; vol. 18(2), pp. 351-361.

[Sc07] Schwartz, J.M. et. al.: Observing metabolic functions at the genome scale. *Genome Biol.*, 2007; vol. 8(6), pp. R123.

[SHK06] Sackmann, A.; Heiner, M.; Koch, I.: Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 2006; vol. 7, pp. 482.

[Th04] Thimm, O. et. al.: MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, 2004; vol. 37(6), pp. 914-939.

[Wu06] Wu, J. et. al.: KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, 2006; vol. 34(Web Server issue), pp. W720-724.