W460–W464 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh476

ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics

Hee-Joon Chung[1], Mingoo Kim[1], Chan Hee Park[1], Jihoon Kim[1] and Ju Han Kim[1,2,*]

[1]Seoul National University Biomedical Informatics (SNUBI) and [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea

Received February 15, 2004; Revised April 18, 2004; Accepted May 3, 2004

ABSTRACT

Biological pathways can provide key information on the organization of biological systems. ArrayXPath (http://www.snubi.org/software/ArrayXPath/) is a web-based service for mapping and visualizing microarray gene-expression data for integrated biological pathway resources using Scalable Vector Graphics (SVG). By integrating major bio-databases and searching pathway resources, ArrayXPath automatically maps different types of identifiers from microarray probes and pathway elements. When one inputs gene-expression clusters, ArrayXPath produces a list of the best matching pathways for each cluster. We applied Fisher's exact test and the false discovery rate (FDR) to evaluate the statistical significance of the association between a cluster and a pathway while correcting the multiple-comparison problem. ArrayXPath produces Javascript-enabled SVGs for web-enabled interactive visualization of pathways integrated with gene-expression profiles.

INTRODUCTION

Cluster analysis is one of the most powerful methods for the exploratory analysis of gene-expression data. Genes clustered on the basis of similarity measures between expression profiles also have positional associations along the chromosomes (1,2), exhibit common *cis*-regulatory elements in their upstream regions (3) and are coordinated by shared sets of regulators (4). For annotation, gene-expression clusters can be assigned to the well-known functional categories of the MIPS classification (5) or the Gene Ontology terms (6) using annotations from public databases (3,7,8).

Biological pathways can provide key information about the organization of biological systems. Major publicly available biological pathway resources, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) (9), GenMAPP (10) and BioCarta (http://www.biocarta.com), provide a large collection of biological pathway diagrams. Tools have been developed to associate microarray gene-expression data with pathway diagrams to create a comprehensive overview and interpretation of the expression profiles (10–13). However, these have some limitations. Some rely on static image files that are difficult to manage. Although some provide dynamically generated diagrams with viewers and editors, the diagrams as well as the tools are encoded in their own proprietary formats. More importantly, some are monolithic, i.e. the data and the visual presentation layers of the pathways are not clearly separated (14). Although Systems Biology Markup Language (SBML) (15) proposes an eXtensible Markup Language (XML)-based standard for encoding pathway data structures, it does not provide one for encoding static and/or dynamic graphics. In JDesigner (http://www.cds.caltech.edu/~hsauro/JDesigner.htm), a Win32 application for editing, visualizing and simulating pathway diagrams encoded in SBML, the visual presentation layer is tightly coupled to the data layer so that it is not easy to integrate them with other useful bioinformatics resources.

One problem that complicates the mapping of expression data onto pathways is that each pathway resource uses different types of identifiers to annotate the nodes containing genes, proteins, enzymes, metabolites, complexes and networks. For example, the identifier of a pathway node may be a gene symbol, GenBank accession number, UniGene ID (sometimes already expired), LocusLink ID, EC (Enzyme Commission) number, or a combination of these, or even null or ambiguous descriptions. Moreover, microarrays of different platforms also use different sets of identifiers for the probes. If one has already mapped the identifiers of predefined pathway nodes and microarray probes, it may be possible to map gene-expression profiles onto pathway diagrams. Without a powerful and general-purpose identifier-mapping engine, however, it may be difficult to map the diverse pathway–node and

*To whom correspondence should be addressed at. Tel: +82 27408320; Fax: +82 7474830; Email: juhan@snu.ac.kr

microarray probe identifiers from different resources reliably. This may in part be why the previously developed tools do not cover all microarray platforms and all public pathway resources but are limited to predefined ones for their analyses.

Clearly separating the data and the presentation layers based on standards and implementing a reliable identifier-mapping engine to map gene-expression profiles correctly onto pathways may benefit pathway-based analyses of microarray data. To our knowledge, there is no web-based service integrating major public pathway resources, mapping diverse identifiers, applying sophisticated statistical tests and visualizing integrated pathway and cluster diagrams. Here we use Scalable Vector Graphics (SVG: http://www.w3.org/TR/SVG), a standard for describing two-dimensional graphics in XML. SVG drawings can be made interactive and dynamic using a supplemental scripting language accessing SVG Document Object Models (DOMs) with a rich set of event handlers. SVG files can then be visualized on the client side with a web browser using a plug-in for SVG.

Here we present ArrayXPath. This is software that (i) receives a clustered gene-expression profile of any microarray platform in a tab-delimited text format via an Internet connection; (ii) automatically resolves the microarray probe identifiers (i.e. GenBank accession number, UniGene ID, LocusLink ID, official gene symbol, SwissProt ID or TrEMBL ID); (iii) searches major public pathway resources (i.e. GenMAPP, KEGG and BioCarta); (iv) maps the different identifier sets between microarray probes and pathway nodes; (v) tests the statistical significance of the association between gene-expression clusters and pathways (hence providing an automated annotation of clusters with the ranked pathways); (vi) visualizes expression levels onto pathways; and (vii) allows web-based user navigation through multiple clusters and pathways enriched with animation features, using Javascript-enabled SVG.

INPUT AND OUTPUT

Input

Input to ArrayXPath is a common tab-delimited text file for a clustered gene-expression profile: ⟨Probe ID⟩-⟨Cluster ID⟩-[⟨Expression level at condition$_i$⟩]. The first column must contain either a GenBank accession number, UniGene ID, LocusLink ID, SwissProt ID, TrEMBL ID or an official gene symbol. The second column contains the cluster ID. The third to i-th columns are optional and contain expression levels. ArrayXPath does not perform cluster analysis *per se*. The input format is designed primarily for a *partitional* clustering algorithm (i.e. K-means or Self-Organizing Maps), but a clustering result from a hierarchical algorithm (i.e. dendrogram) may be applied by choosing a threshold carefully.

Output

ArrayXPath produces a list of the best matching pathways for each cluster with statistical significance scores of non-random association. Relevant pathways are listed in ascending order of P-values (and multiple-comparison corrected Q-values) (16). ArrayXPath provides a summary statistic for the overall mapping between input clusters and all pathways matched. If one chooses a pathway from the list, ArrayXPath outputs a Javascript-enabled SVG file, color-coded both by expression level and by cluster membership at each pathway–node level.

One can zoom in and out for better navigation through multiple clusters and pathways with animation features using Javascript-enabled SVG. One can choose particular experimental conditions and any combination of clusters. The sequential alteration of expression levels across conditions can be viewed as an animated visual. Because it is tricky to overlap more than one set of expression levels (i.e. more than one condition) at a time on the same pathway diagram, ArrayXPath provides an additional plot diagram to show the expression levels of all clusters across all conditions (Figure 1).

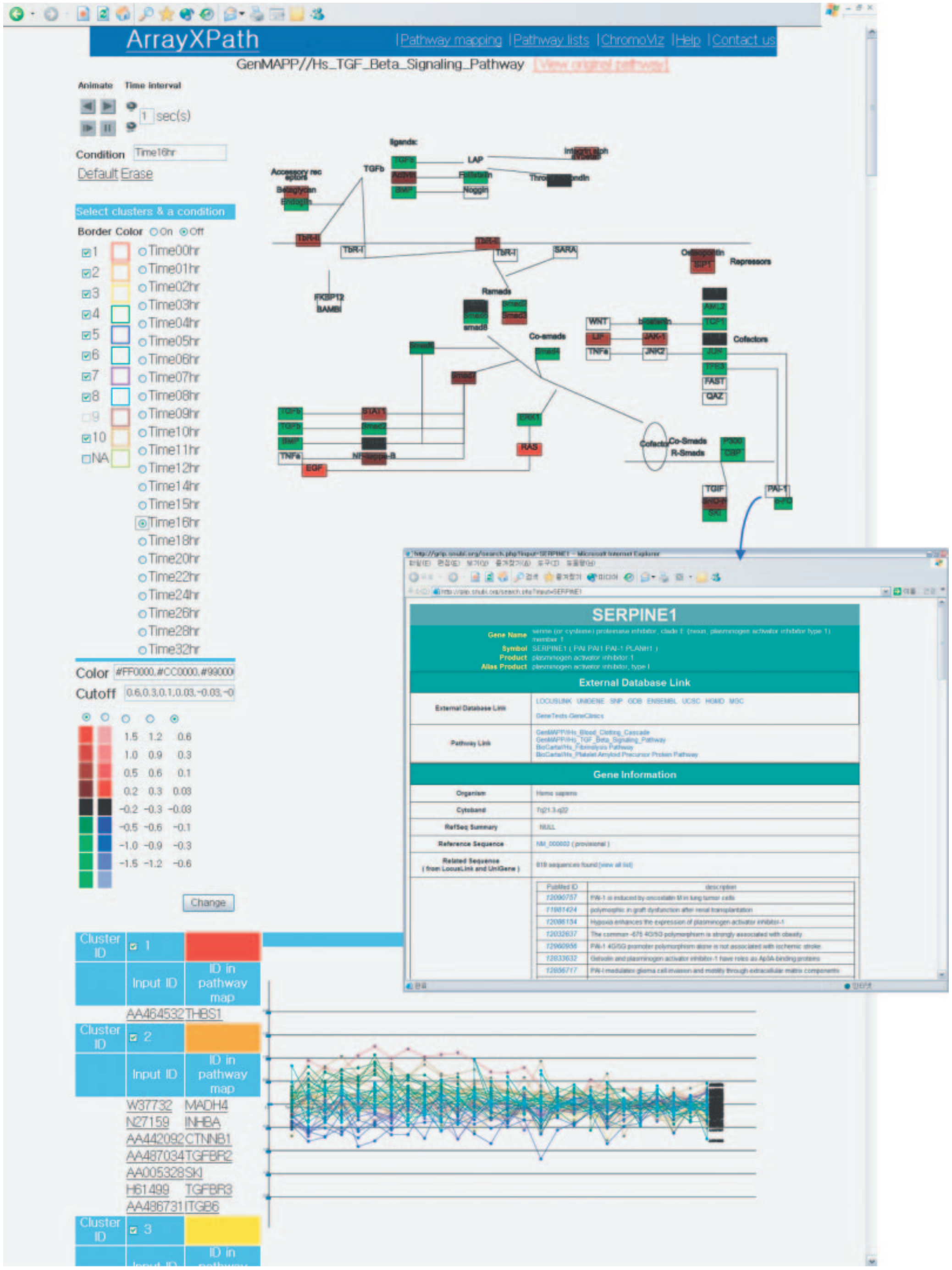
A node in a pathway may be composed of more than one element. For example, an enzyme complex in a KEGG pathway may be composed of many proteins. ArrayXPath automatically resolves the appropriate element IDs for the mapping process with microarray probes, inserts the elements as a legend and separately visualizes the expression levels at each element level. Each pathway node is enriched with a hyperlink to an automated annotation page for the corresponding gene product(s) provided by our integrated database, GRIP (Genome Research Informatics Pipeline).

METHODS

Pathway integration and resolving diverse identifiers

ArrayXPath searches publicly available major pathway resources including KEGG, GenMAPP and BioCarta. We have created a repository of meta-information by parsing SBML files for KEGG and HTML files for GenMAPP (http://www.genmapp.org/MAPPSet-Human/MAPP_index.htm) and BioCarta (http://www.biocarta.com/genes/allPathways.asp). As mentioned in the Introduction, a variety of identifiers, including GenBank accession number, UniGene ID, LocusLink ID, EC number, official gene symbol, SwissProt ID and TrEMBL ID, are inconsistently used for the pathway nodes as well as microarray probes, resulting in enormous ambiguity in integrating data from different resources. We have successfully integrated the major databases including GenBank, UniGene, LocusLink, SwissProt, Ensemble and UCSC Golden Path: refGene, knownGene, all_mRNA and all_est. Homologous pairs for the probes are resolved using NCBI's Homologene. NetAffyx is used for Affymetrix oligonucleotide arrays (http://www.affymetrix.com/analysis/index.affx). When one inputs a clustered gene-expression profile complying with the ArrayXPath input format (see Input and Output), ArrayXPath automatically matches the probe identifiers of microarray data to the identifiers of pathway nodes using a pre-computed table of identifiers from the major databases. When a pathway node is a composite type, i.e. consists of more than one element, ArrayXPath separately matches and visualizes each probe identifier to the corresponding individual element of the composite object.

Table 1 shows the distribution of the pathway nodes identified from KEGG, GenMAPP and BioCarta for *Homo sapiens*. We found 1942 redundant nodes representing genes and proteins for the 45 GenMAPP pathways. Among the



Condition Time16hr

1 sec(s)

Border Color ○On ●Off

☑1		○Time00hr
☑2		○Time01hr
☑3		○Time02hr
☑4		○Time03hr
☑5		○Time04hr
☑6		○Time05hr
☑7		○Time06hr
☑8		○Time07hr
□9		○Time08hr
☑10		○Time09hr
□NA		○Time10hr
		○Time11hr
		○Time12hr
		○Time14hr
		○Time15hr
		●Time16hr
		○Time18hr
		○Time20hr
		○Time22hr
		○Time24hr
		○Time26hr
		○Time28hr
		○Time32hr

Color #FF0000,#CC0000,#99000

Cutoff 0.6,0.3,0.1,0.03,-0.03,-0

1.5	1.2	0.6
1.0	0.9	0.3
0.5	0.6	0.1
0.2	0.3	0.03
-0.2	-0.3	-0.03
-0.5	-0.6	-0.1
-1.0	-0.9	-0.3
-1.5	-1.2	-0.6

Cluster ID	Input ID	ID in pathway map
☑ 1	AA464532	THBS1
☑ 2	W37732	MADH4
	N27159	INHBA
	AA442092	CTNNB1
	AA487034	TGFBR2
	AA005328	SKI
	H61499	TGFBR3
	AA486731	ITGB6
☑ 3		

http://grip.snubi.org/search.php?input=SERPINE1 – Microsoft Internet Explorer

SERPINE1

Gene Name serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
Symbol SERPINE1 (PAI PAI1 PAI-1 PLANH1)
Product plasminogen activator inhibitor 1
Alias Product plasminogen activator inhibitor, type I

External Database Link	
External Database Link	LOCUSLINK UNIGENE SNP GDB ENSEMBL UCSC HGMD MGC GeneTests GeneClinics
Pathway Link	GenMAPP/Hs_Blood_Clotting_Cascade GenMAPP/Hs_TGF_Beta_Signaling_Pathway BioCarta/Hs_Fibrinolysis Pathway BioCarta/Hs_Platelet Amyloid Precursor Protein Pathway

Gene Information	
Organism	Homo sapiens
Cytoband	7q21.3-q22
RefSeq Summary	NULL
Reference Sequence	NM_000602 (provisional)
Related Sequence (from LocusLink and UniGene)	819 sequences found [view all list]

PubMed ID	description
12090757	PAI-1 is induced by oncostatin M in lung tumor cells
11981424	polymorphic in graft dysfunction after renal transplantation
12086134	Hypoxia enhances the expression of plasminogen activator inhibitor-1
12032637	The common -675 4G/5G polymorphism is strongly associated with obesity
12960956	PAI-1 4G/5G promoter polymorphism alone is not associated with ischemic stroke
12833632	Gelsolin and plasminogen activator inhibitor-1 have roles as Ap3A-binding proteins
12856717	PAI-I modulates glioma cell invasion and motility through extracellular matrix components

Figure 1. ArrayXPath maps gene-expression profiles onto pathway resources and visualizes pathway diagrams color-coded by gene-expression level. The control panel on the left side permits users to interactively navigate the pathway and plot diagrams. As shown in the inset screenshot, ArrayXPath provides summary information for each pathway node by integrating major public databases.

1454 non-redundant elements, ArrayXPath successfully assigned 1391 gene products (95.7%) to official gene symbols ($n = 1329$; 91.4%), LocusLink IDs ($n = 39$; 2.7%) or SwissProt IDs ($n = 23$; 1.6%). Only 63 (4.3%) remain unresolved because of ambiguity. KEGG has 256 non-composite (i.e. simple) and 121 composite elements (i.e. enzymes), containing 256 and 505 gene products, respectively. Among the 256 simple-type elements, 21 appear as members of composite-type elements. Overall, KEGG has 740 unique elements and ArrayXPath successfully assigned all of them (100%) either to official gene symbols ($n = 720$; 97.3%) or to LocusLink IDs ($n = 20$; 2.7%).

Overall, ArrayXPath identified 3008 gene products for the three major pathways. We created a pre-computed table linking these elements to all resolvable GenBank accession numbers, to the UniGene, LocusLink, SwissProt and TrEMBL IDs

Table 1. Distribution of pathway–node identifiers among the major pathway resources

Pathway	No.	Gene/protein			ID resolution					Metabolite	Embedded pathway	Free-text description
		Simple	Composite	Redundant	Total	OGS	LL	SP	UR			
KEGG	70	(256)[a] (505)[a] 740	(121)	(637) (469) (1106)	740	720	20	0	0	1896 (2624)	0 (0)	121 (275)
GenMAPP	45	1454		(1942)	1391	1329	39	23	63	83 (97)	4 (4)	130 (372)
BioCarta	346	1584		(8976)	1584	1580	4	0	0	0 (0)	50 (141)	18 (53)
Overall	461	3008		(12706)	2945	2859	63	23	63	1979 (2721)	54 (145)	55 (146)

[a] Among the simple ($n = 256$) and composite ($n = 505$) KEGG elements, the number of redundant elements was 21 such that we identified 740 (= 256 + 505 − 21) unique elements in the 70 KEGG pathways. Numbers in parentheses are redundant counts. For example, KEGG has 121 composite elements containing 505 identifiable gene products, 21 of which are redundant.

OGS, official gene symbol; LL, LocusLink; SP, SwissProt; UR, unresolved.

and to official gene symbols for the reliable mapping of incoming microarray probe identifiers. The accuracy, reliability and coverage of this identifier-resolving process are essential for the validity of the pathway lists, which are thus matched to a cluster of microarray probes.

Statistical significance testing and false discovery rates

ArrayXPath determines the statistical significance of the association between a gene-expression cluster and a pathway in terms of the non-random proportion of matched entities. Specifically, ArrayXPath applies Fisher's exact test by constructing a 2 × 2 contingency table containing the two cluster memberships (within and without the cluster) as row variables and the pathway memberships (within and without the pathway) as column variables. We used Fisher's exact test because a large sample approximation is inappropriate in the pathway case (a 2 × 2 table often contains a cell with expected values <5).

Because this approach simultaneously tests the statistical significances of the associations of a cluster to multiple pathways, we also have to deal with the problem of 'multiple hypothesis testing'. Two types of error measurements are commonly used in multiple hypothesis testing: the family-wise error rate (FWER) and the false discovery rate (FDR). The FWER offers a very strict error measure of at least one false positive result among all significant hypotheses. The FDR is defined as the expected proportion of false positive results among all rejected hypotheses multiplied by the probability of making at least one rejection (17). The FDR offers a much less strict criterion and hence leads to an increase in power. Pathways scores (P-values) smaller than a threshold are considered potentially significant. The percentage of such pathways identified by chance is the false discovery rate. The P-value cut-off was decided by determining the FDR following the scheme of Storey *et al.* (16).

The FDR provides a useful measure of the overall accuracy of a set of significant matches. Because our interest lies in ranking each pathway according to a certain threshold value, the Q-value is preferred as a measure of significance for any individual pathway. Assuming that null P-values after the Fisher's exact test are uniformly distributed in the density plot of P-values, the proportion of truly null matches (those that are equal to π_0) can be estimated as the height of the flat proportion of P exceeding a certain threshold: for the detailed algorithm for fine-tuning π_0 we refer the reader to Storey *et al.* (16). After the Q-value is calculated, the maximum P-value with an estimated Q-value less than or equal to the given threshold is chosen as the P-value cut-off. We calculated the Q-values for the matched pathways for each cluster and listed them in ascending order.

DISCUSSION

ArrayXPath is a web-based service for mapping and visualizing microarray gene-expression profiles onto major biological pathway resources. It permits one to input a clustered gene-expression profile in a tab-delimited text format. Standard representation of pathway diagrams and the clear separation of the data and the visual presentation layers will benefit flexible and extensible integration of bioinformatics modules as well as heterogeneous genomic data.

We found ~3000 non-redundant genes and proteins in the three major pathway resources for *H.sapiens*, which is a relatively small number compared with the estimated number of genes for our species. Once we have integrated public pathway resources in a standard manner, it may be possible to extend the pathway resources to biomedical literature and factual biodatabases, using text mining and machine learning techniques. Deciphering the crosstalk among pathways may also be an interesting issue for the system-level understanding of life. Standard web-based integration of a wide range of bioinformatics resources will obviously help advance biological science.

ACKNOWLEDGEMENTS

The authors thank all SNUBI members and external testers for their effort in inputting and evaluating the pathway diagrams. This study was supported by a grant from the Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (03-PJ10-PG13-GD01-0002).

REFERENCES

1. Roy,P.J., Stuart,J.M., Lurd,J. and Kim,S.K. (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature*, **418**, 975–979.

2. Lercher,M.J., Urrutia,A. and Hurst,L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet*., **31**, 180–183.
3. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet*., **22**, 281–285.
4. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet*., **34**, 166–176.
5. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*., **30**, 31–34.
6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
7. Dennis,G.,Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*., **4**, R60.
8. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.
9. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
10. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet*., **31**, 19–20.
11. Grosu,P., Townsend,J.P., Hartl,D.L. and Cavalieri,D. (2002) Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res*., **12**, 1121–1126.
12. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.
13. 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.
14. Pressman,R.S. (1997) *Software Engineering: A Practitioner's Approach 4th edn*. McGraw-Hill, New York, NY, pp. 341–361.
15. Finney,A. and Hucka,M. (2003) Systems biology markup language: Level 2 and beyond. *Biochem. Soc. Trans*., **31**, 1472–1473.
16. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*., **100**, 9440–9445.
17. 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.