BIOINFORMATICS
Vol. 20 no. 4 2004, pages 507–517
DOI: 10.1093/bioinformatics/btg436

Tailored gene array databases: applications in mechanistic toxicology

Tatiana V. Karpinets[1], Brent D. Foy[1,*] and John M. Frazier[2]

[1]Department of Physics, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA and [2]Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433-7400, USA

Received March 19, 2003; revised July 16, 2003; accepted August 9, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: The development of an annotated global database suitable for a wide range of investigations is a challenging and labor-intensive task. Thus, the development of databases tailored for specific applications remains necessary. For example, in the field of toxicology, no annotated gene array databases are now available that may assist in the correlation of changes in gene activity to cellular functions and processes associated with the toxic response.

Results: As an example of a tailored annotated database, an attempt was made to systematize available biological information on genes present on the Affymetrix Rat Toxicology U34 GeneChip, with a focus on how the gene products relate to liver cells and their response to chemical toxins. The information collected was imbedded in a local relational database to analyze data obtained in toxicological gene array experiments with hydrazine–exposed hepatocytes. The advantages and benefits of the tailored database in the biological interpretation of the results are demonstrated.

Availability: The database and gene expression data set are available at http://www.wright.edu/~tatiana.karpinets/

Contact: brent.foy@wright.edu

INTRODUCTION

Gene expression analysis opens new perspectives to reveal biological mechanisms that underlie different patterns of genome activity and to simulate these mechanisms on a qualitative basis. However, this approach has not been fully exploited because of the challenging task of biological interpretation of the results obtained from the quantitative estimation of expression of thousands of genes.

One approach to seek biological meaning has been to cluster mathematically the expression data according to their expression intensities and/or their dynamic patterns (Eisen *et al*., 1998; Li *et al*., 2002; Kerr and Churchill, 2001). The clustering of genes with similar expression-versus-time patterns is based on the hypothesis that genes that share a cluster may be associated with the same causal mechanism or pathway (DeRisi *et al*., 1997; Chu *et al*., 1998; Chang *et al*., 2001). The identification of the biological mechanism that underlies any large grouping of genes is a time-consuming task for researchers. It involves not only the analysis and the selection of information about each gene product in clusters, but also the organization and presentation of this information.

The development of tools for automatic accomplishment of this task is beginning. A method and corresponding database was developed that use keywords from the published literature linked to specific genes to estimate similarity of genes within a cluster of genes (Masys *et al*., 2001). The *Arabidopsis thaliana* Database project aims to transform the database into an integrated biological database resource by integrating diverse data, tools, queries and visualization capabilities (Schoof *et al*., 2002). The National Center for Toxicogenomics is developing the first public toxicogenomics knowledge base that combines molecular expression data sets with metabolic, toxicological pathway, and gene regulatory network information relevant to environmental toxicology and human disease (Waters *et al*., 2003).

The Gene Ontology (GO) tool is also making progress along these lines (Ashburner *et al*., 2000; Xie *et al*., 2002). GO consists of three independent ontologies: (1) biological processes, (2) molecular function and (3) cellular component. Each of these ontologies is based on parent–child relationship, i.e. a child term can be an instance of its parent or its part. A gene product may have several molecular functions, take part in several biological processes and be associated with several cellular components. GO provides a set of structured vocabularies to describe gene products in any organism. The *Saccharomyces* Genome Database (SGD), the FlyBase database of the *Drosophila* genome and the database of Functional Annotation of RIKEN Mouse cDNA Clones (FANTOM DB) annotate genes using GO (Dwight *et al*., 2002; The FlyBase Consortium, 2002; Bono *et al*., 2002). Document classification methods may be applied to associate GO codes to literature abstracts and thus to the genes associated with the abstracts (Raychaudhuri *et al*., 2002). A computational

*To whom correspondence should be addressed.

platform for GO annotations, and analysis of the resultant GO annotations of human proteins has also been recently reported (Xie *et al*., 2002).

The implementation of the GO approach in specific studies, however, has some limitations. The primary issue is that the GO is constructed to be a universal categorization of genes and gene products across species and states. However, when attempting to use it for a specific state and species, the universal categories are not always the same ones that experts in the sub-field might choose. Another difficulty is the absence of pathways categories in GO. This makes it difficult to present gene products in the context of interactions between them. An additional problem is that the nested structure of GO (most terms have descendants, which in turn have other descendants and so on), means that modifying a category to incorporate a new task is difficult. The implementation of GO in toxicological studies is also limited by the fact that this universal vocabulary annotates gene products with terms reflecting their normal activity and location. Unique functions, processes and localizations observed in mutant or disease state are not usually included.

This paper describes an easy to use relational database designed for the Affymetrix Rat Toxicology U34 (RT U34) Array and tailored to the specific task of evaluation and characterization of biological mechanisms relating to adverse responses of hepatocytes to chemical toxicants. This database incorporates structural information about cellular functions, pathways and localizations for all gene products represented in Rat Tox U34 GeneChip.

SYSTEMS AND METHODS

Rat toxicology U34 array

The RT U34 GeneChip contains 1031 probes representing more than 850 genes and expressed sequence tag (EST) clusters selected from the UniGene database. The array includes genes that encode proteins involved in the basic metabolic processes, antioxidation, detoxification, immune response, signal transduction, cell proliferation and apoptosis. Most of these gene products are especially active in liver cells. Hepatocytes retain their characteristic in cell culture and therefore have proved valuable in toxicological studies (Guillouzo, 1998).

Sources and types of information

Information about genes present on the Rat Tox U34 GeneChip were collected manually: (1) from scientific papers that were published during the last 5 years that have abstracts presented at PUBMED, STKE, MEDLINE or ChemPort; and (2) from Internet databases including GenBank (Benson *et al*., 2002), Kyoto Encyclopedia of Genes and Genomes (Kanehisa *et al*., 2002), Cell Signaling Networks Database (Takai-Igarashi and Kaminuma, 1999), Database of Interacting Proteins (Xenarios *et al*., 2002), Enzyme Database BRENDA (Schomburg *et al*., 2002). If sufficient information was available concerning a particular gene product, priority was given to studies related to hepatocytes and stress responses. The collected information included the following items:

(1) Gene systematic name (Affymetrix probe name) and GenBank accession number.

(2) Gene description supplied by Affymetrix.

(3) Name of family or subfamily to which the gene product belongs.

(4) Cellular functions of the gene product in hepatocytes and a short description of those functions.

(5) Cellular localizations of the gene product in liver cells.

(6) Pathways and their complexes in liver cell, in which the gene product may be involved in the normal state of the cell and/or under stress conditions.

(7) References, with hyperlink, to sources where this information was obtained.

Database development

This information was analyzed to identify entities within classes (collections of basic objects that represent characteristics of genes and their products), their attributes and entities, which are essential to characterize cellular responses to stress condition and metabolism of xenobiotics. An entity-relationship model of the database was developed according to the identified entity classes, then implemented as a relational database in Microsoft Access, and populated with the collected information on gene products. To automate the biological interpretation of genes in a cluster (we will refer to list of genes in a cluster as a Gene set), we identify the relationships between the information in the database and each gene in the Gene set and provided a quick access to it through database queries.

Experimental data for the demonstration of database application

Experimental data from hydrazine treatment of rat liver cells (unpublished data) serve to demonstrate the use of the database in the extraction of biologically meaningful information from the gene expression data. Hydrazine is a highly toxic chemical used as a propellant. The involvement of this toxicant in DNA damage, oxidative stress, apoptosis and cancerogenic transformations of hepatocytes is well established (Keller, 1988; Sax, 1990; Kutsenko *et al*., 2001; Hussain and Frazier, 2002, 2003). The experimental protocol involved a time course study of three experimental groups distinguished by the exposure of rat hepatocytes with three concentration of hydrazine: 0 (control), 50 and 75 mM for 2 h. We will refer to these groups as Time-Matched Control, Group-50 and Group-75. Measurement of gene expression was performed just before the 2 h treatment period (referred to as the Pretreatment Control) and at 0, 3, 6, 12 and 24 h post treatment. Two replicate GeneChips were hybridized at each time point. The level of gene expression in the Pretreatment Control was considered as the initial value for gene expressions and used to calculate the log-ratios of gene expressions for the rest of the time points.

Three clusters of genes were obtained by statistical processing of the results. The first cluster (referred to as the G-cluster) contained 311 genes that exhibited significant changes in expression at any time point as a result of hydrazine treatment (Group-50, Group-75 or both) when compared with the Time-Matched Control. The selection of genes in this cluster was implemented by a comparison of log-ratios of gene expressions in Group-50 and Group-75 with Time-Matched Control for each time point by means of a Student's t-test ($P \leq 0.05$). A single significant difference at any time point or exposure concentration was sufficient to include a gene in G-cluster. This is a relatively relaxed criteria designed to ensure inclusion of the majority of those genes that may be affected by hydrazine treatment. Genes in this G-cluster were then divided into two clusters according to the sign and value of the correlation coefficients calculated for the level of gene expression versus time. The second cluster, called the C-cluster, is a subset of the G-cluster and contains 173 genes that have a common positive (>0.30) or negative (<-0.30) correlation coefficient for all three experimental groups. The low value of 0.3 was chosen for the correlation coefficient because the intent was to group genes that have a general upward or downward trend for all treatments, rather than to select genes, which have a strong linear correlation between expression and time. The remaining 138 genes of the G-cluster were assigned to the D-cluster. Thus, the D-cluster contained those genes from the G-cluster in which at least one of the three experimental groups exhibited a time course significantly different from the other two groups.

IMPLEMENTATION

Database model

To elucidate biological information contained within clusters of genes obtained by mathematical analysis of gene array data, the database should categorize gene products according to: (1) their function in stress responses; (2) their involvement in basic pathways or group of pathways related to metabolism of xenobiotics, carcinogenicity, different types of subcellular damage and cellular pathologies and (3) the localization of gene products with respect to cellular compartments. To provide the necessary annotation for inferences in the database, we identified eight main data classes into which the collected information was categorized:

Genes: The class Genes is populated with entities that identify the probe sets on the gene array. In the database described here, this class includes entities that map to the probe sets present on the RT U34 GeneChip (number of entities = 1031). We will refer to these entities as genes.

Prototypes: This class consists a list of entities, each of which identifies probe sets whose associated genes are categorized identically. The redundancy of probes on the array for the same gene and lack of complete biological knowledge about all genes make the introduction of this class very helpful in avoiding repetitive input of information about entities in the class Genes. In other words, all the genes in a single entity of the class Prototypes have the exact same entity descriptors for the three database classes (see below for definitions): Functions, Pathways and Localizations. The concept of a Prototypes class is similar to the conventional definitions of protein families or superfamilies. However, in comparison to the class Prototypes, the criteria to cluster genes into a family or superfamily are more rigorous. They imply not only similar function, but also having common evolutionary origins and structures (Lo Conte *et al*., 2000). For the present, the class Prototypes includes 234 entities, e.g. interleukin 4, aldehyde reductase, wee1, tyrosine kinase. If one gene in a set of genes that constitute a particular Prototypes entity is discovered to have a key feature distinct from other members in that Prototypes entity, then it is broken out into a new entity of its own. Several entities in the class Genes may belong to the same entity in the class Prototypes, but only one entity from the class Prototypes may characterize each entity from the class Genes.

Functions: The class Functions is a collection of entities that provide a simplified categorization of gene products according to their cellular functions, with an emphasis on adverse stress responses. Entities in this class (Table 1) associate each gene product with a more descriptive name than the class Prototypes entity name, and in this way integrate biological information about the functions of 234 Prototypes entities into 26 entities of the class Functions. In fact, this class provides the next level of compressed biological information (after the class Prototypes) associated with molecular functions of gene products in stress responses, detoxification of xenobiotics and disease states. Most probes of the array represent enzymes related to these processes and entities of this class specify their stress-related function, like antioxidant, DNA repair enzyme and so on. A large group of probes represents cellular functions of proteins involved in a series of steps (cytokines → receptors → MAP kinases and their regulators, secondary messenger molecules and their regulators → transcription factors) that underlie stress responses, particularly the regulation of cell growth, proliferation and differentiation as a result of adverse stimuli. Several important entities in class Functions, though not populated with genes on the microarray, are represented by regulators of the cell cycle, apoptosis and inflammation. Other entities are structural components and transporters. Several entities of the class Prototypes may

Table 1. Entities in the classes Functions Group and Functions

Entities of the Functions Group class	Entities of the Function class	Number of probes
Enzymes related to the detoxification of xenobiotics, DNA repair and disease state (393)	Antioxidants	20
	Cytochromes 450	77
	Conjugating enzymes	63
	Other enzymes involved in detoxification of xenobiotics	29
	DNA repair enzymes	33
	Other enzymes related to metabolism	171
The regulators of signal transduction in stress responses (242)	Chaperones	51
	Cytokines	52
	Transmembrane receptors	34
	Mitogen-activated protein kinases	28
	MAP kinase phosphatases	4
	Secondary messenger molecules	17
	Regulators of G-protein signaling	2
	Transcriptional factors	46
Cell cycle regulators (21)	Transcription factor regulators	8
	Cyclin-dependent kinases	16
	Cyclin-dependent kinase inhibitors	5
Apoptosis regulators (35)	Caspases	7
	Apoptosis activators	10
	Apoptosis inhibitors	18
Regulators of inflammatory response (7)	Adhesion molecules	4
	Acute phase proteins	3
Other entities (57)	Responsive elements	1
	Structural components	37
	Transporters	19
Unknown (276)	Unknown ESTs	276
Total		1031

The number of probes on the RT U34 GeneChip assigned to each entity is indicated.

have the same entity in the class Functions, and several entities in the class Functions may belong to a single entity of the class Prototypes. In the current database, all identified entities in the class Prototypes have only one well-documented cellular function directly related to stress responses.

Functions Group: The class Functions Group segregates entities of the class Functions into broader categories (Table 1) with related biological activities. This higher level organization of our knowledge about functions allows for the integration of biological information about the functions of gene products represented by probes on the array.

Pathways: The class Pathways (Table 2) is an attempt to capture the interactions and/or connections between gene products and compress this information into 48 known metabolic and regulatory pathways or cellular processes. This class provides a categorization of pathways and cellular processes that may be involved in stressful responses and disease state or may be impaired as a result of them. Entities in the class represent 25 metabolic pathways involved in regular metabolic processes and biodegradation of xenobiotics, and 11 regulatory pathways involved in cell growth and death, replication and repair, signal transduction and cell communication. We also specified 12 cellular processes in this class implicated in immune response, antioxidation and structure of cellular component. Many of these pathways are presented in the Kyoto Encyclopedia of Genes and Genomes (Kanehisa *et al*., 2002) and on the website of Signal Transduction Knowledge Environment (Gough and Ray, 2002). Examples of entities in the class are steroid hormone metabolism, p53 pathway and phase II metabolism of xenobiotics. As the implication of gene products in the cellular processes will have different levels of specification, entities in this class represent both traditional pathways and more general cellular processes, such as the ras/raf/MEK/ERK pathway and tumorigenic transformations. There is a 'many-to-many' type of relationship between entities in the class Prototypes and entities in the class Pathways. Analysis of the collected information shows that in comparison with the Functions class, most gene products under consideration are implicated in more than one pathway.

Table 2. Entities of the classes Pathways Group and Pathways

Entities of the Pathways Group class	Entities of the Pathways class	Number of probes
Basic metabolic pathways	Hem synthesis; growth hormone metabolism; Ca homeostasis and Ca signaling; glycolytic pathway; gonadogenesis; cholesterol metabolism; formation of the extracellular matrix; catecholamine metabolism; respiratory chain; lipid metabolism; membrane transport; vesicular trafficking; urea cycle; trafficking between nucleus and cytoplasm; protein and amino acid metabolism; purine, pyrimidine and pyridine metabolism; iron homeostasis; sulfur metabolism, melatonin synthesis; steroid hormone metabolism	243
Biodegradation of xenobiotics	Phase II metabolism of xenobiotics; phase I metabolism of xenobiotics; other processes related to detoxification; ubiquitin pathway	206
Cell growth and death	Protein repair and chaperoning; tumorigenic transformations; angiogenesis; apoptosis; necrosis; ras/raf/MEK/ERK pathway; p38 pathway; p53 pathway JNK/SAPK pathway	422
Signal transduction and cell communication	Signal transduction; cell communication	91
DNA replication and repair	DNA replication; DNA repair	48
Immune and inflammatory responses	Immune response; inflammatory response	121
Antioxidation	Glutathione metabolism; oxidative metabolism of xenobiotics; peroxisome proliferation	151
Structural components	Cell junctions; cytoskeleton; proteasome structure; peroxisome structure; ribosome structure	38
Unknown	Unknown ESTs	276

The number of probes on the RT U34 GeneChip assigned to each entity is indicated.

Pathways Group: The class Pathways Group separates entities of the class Pathways into larger-categories based on similar properties relating to biological processes (Table 2). Entities in the class Pathways Group provide logical connections between biological processes presented by this class to stress response.

Localizations: Entities of the class Localizations (Table 3) identify cellular and extracellular locations where gene products perform their molecular functions. Most gene products under consideration may be found in several cellular components. Similar to the class Pathways, entities in the class Localizations may relate to different levels of biological organization depending on the available information about gene products.

References: This class is the database used to identify references to websites and papers with hyperlinks, where the information about the genes on the microarray and their encoded proteins were obtained. Currently, it includes 302 entities.

Table 3. Entities of the class Localizations

Entities of the Localizations class	Number of probes
Cytoplasm	269
Mitochondria	196
Nuclear	143
Endoplasmic reticulum	102
Plasma membrane	90
Extracellular matrix	59
Peroxisome	20
Cytosceleton	11
Gap junctions	11
Proteasome	10
Golgi apparatus	6
Ribosome	4
Centrosome	1
Unknown ESTs	276
No info	59

The number of probes on the RT U34 GeneChip assigned to each entity is indicated.

Within the database, additional information relating to the annotation of the genes is stored. This information includes a short gene description, family name for the encoded protein, a short description of its molecular function, a full text description of the protein that includes its molecular functions in liver cells and the biochemical mechanism by which these functions are implemented (if available), and possible interactions with other proteins (if available). The database also provides helpful hyperlinks for all probes present in the array to GenBank, where full nucleotide sequences, citations and other available information about the gene may be extracted and to recent scientific papers in the published literature where comprehensive information on the gene product may be obtained.

Queries to the database

Queries to the database provide comprehensive information about any collection of genes given as a list of their Affymetrix names or GenBank accession numbers and are presented by means of tables or histograms. Typically a cluster of genes is generated using a pattern-recognition or other mathematical clustering algorithm, and then queries to the database reveal (i) the number of genes from this cluster involved in different cellular functions, pathways and locations and (ii) the representation of these cellular functions, pathways and locations. The latter parameter characterizes the extent to which an entity is over- or under-represented in a cluster. The underlying assumption is that if the selection of genes in a cluster is random, then the percentage of genes of the entity in the entire array will equal the percentage of genes of the entity in the cluster. We characterize the representation of a j-th entity of a i-th class in the cluster by means of the representation index (R_{ij}). This index is calculated as

$$R_{ij} = \frac{\left|PC_{ij} - PA_{ij}\right|}{PCE_{ij} - PA_{ij}}$$

where PC_{ij} is the percentage of genes in the j-th entity of the i-th class (referred to as i, j-genes) in the cluster calculated as $PC_{ij} = m/M$ (m is the number of i, j-genes in the cluster and M is the total number of genes in the cluster). PA_{ij} is the percentage of i, j-genes on the microarray calculated as $PA_{ij} = n/N$ (n is the number of i, j-genes on the microarray and N is the total number of genes on the microarray (1031 for the RT U34 array). PCE_{ij} in the denominator is the extreme value for the percentage of i, j-genes in the cluster and depends on whether the i, j-entity is over- or under-represented in the cluster. When the entity is over-represented in the cluster, $PC_{ij} > PA_{ij}$, then PCE_{ij} equals the maximum possible value for the percentage of genes in the cluster, i.e. $PCE_{ij} = a/M$, where a is the smaller of n or M. If the total number of genes in the cluster is less than the number of i, j-genes on the microarray, then $a = M$ and $PCE_{ij} = 1$. On the other hand, if the total number of genes in the cluster is greater than the number of i, j-genes on the array, then $a = n$ and $PCE_{ij} = n/M$. When $PC_{ij} < PA_{ij}$ then PCE_{ij} equals the minimal value of the percentage of i, j-genes in the cluster, i.e. $PCE_{ij} = 0$. In words, the equation above calculates the magnitude of the actual change in the percentage of genes for an entity divided by the maximum possible change of the percentage of genes for that entity, where the maximum possible change is in the same direction (positive or negative) as the actual change. Thus, R_{ij} has the same dynamic range among different entities and different size clusters. If the representation index (R_{ij}) is negative then the j-th entity of the i-th class is under-represented in the cluster, and if it is positive, the entity is over-represented. The R_{ij} value is 0 if the representation of the j-th entity of the i-th class in the cluster is the same as on the array ($PC_{ij} = PA_{ij}$). R_{ij} equals $+1.0$ if the cluster contains the maximum number of i, j-genes as possible for that cluster, and R_{ij} equals -1.0 if the entity is not represented in the cluster at all ($PC_{ij} = 0$).

In addition to analyzing clusters that were selected by some *ad hoc* mathematical technique, the database may be used to perform an initial clustering of the genes to select genes having a particular biological similarity, such as a certain pathway or function. Then mathematical methods may be applied to estimate quantitatively the changes in gene expression in these biologically similar clusters. Both approaches may compliment each other in a study; however, the first application is demonstrated in the example below.

An example of database implementation to the analysis of experimental data

The objective of this analysis was to identify changes in cellular functions and processes that characterize the response of hepatocytes to hydrazine. It was implemented by the following sequence of steps: (i) the categorization of genes in three clusters as was described in the Methods section; (ii) the identification of over- or under-represented entities of the class Functions and of the class Pathways within the clusters via queries to the database and (iii) a *post hoc* analysis of changes in the expression of genes that constitute the over-represented entities in each cluster, including a detailed look at whether gene expression was increasing or decreasing with time.

The G-, C- and D-clusters were analyzed using the Pathways and Functions class entities separately. The representation indices for entities in the Pathways class is presented in Figure 1 (data for the Functions class analysis are not shown). By inspecting the representation indexes of the Pathways class entities in the clusters, we identified biological processes in hepatocytes that are over-represented in the C-cluster and therefore common to the stress response of hepatocytes in culture regardless of the presence or absence of hydrazine; and in the D-cluster, and therefore specific to their treatment with one or both exposure concentrations of hydrazine. Gene products identified in the over-represented entities of the Functions and Pathways class analyses within the D-cluster, and trends in their gene expressions are given in Table 4.

A large group of pathways in the Pathways class is composed of entities that are over-represented ($R > 0$) in the G-cluster and the C-cluster, but under-represented ($R < 0$) in the D-cluster (Fig. 1). This means that changes in these processes are not specific cellular responses following hydrazine treatment. They include many regular metabolic and detoxification pathways, e.g. cholesterol metabolism, lipid metabolism, steroid hormone metabolism, structural components of cytoskeleton, phase I and II metabolism of xenobiotics and processes related to the function of peroxisomes. Most genes in these entities have decreasing

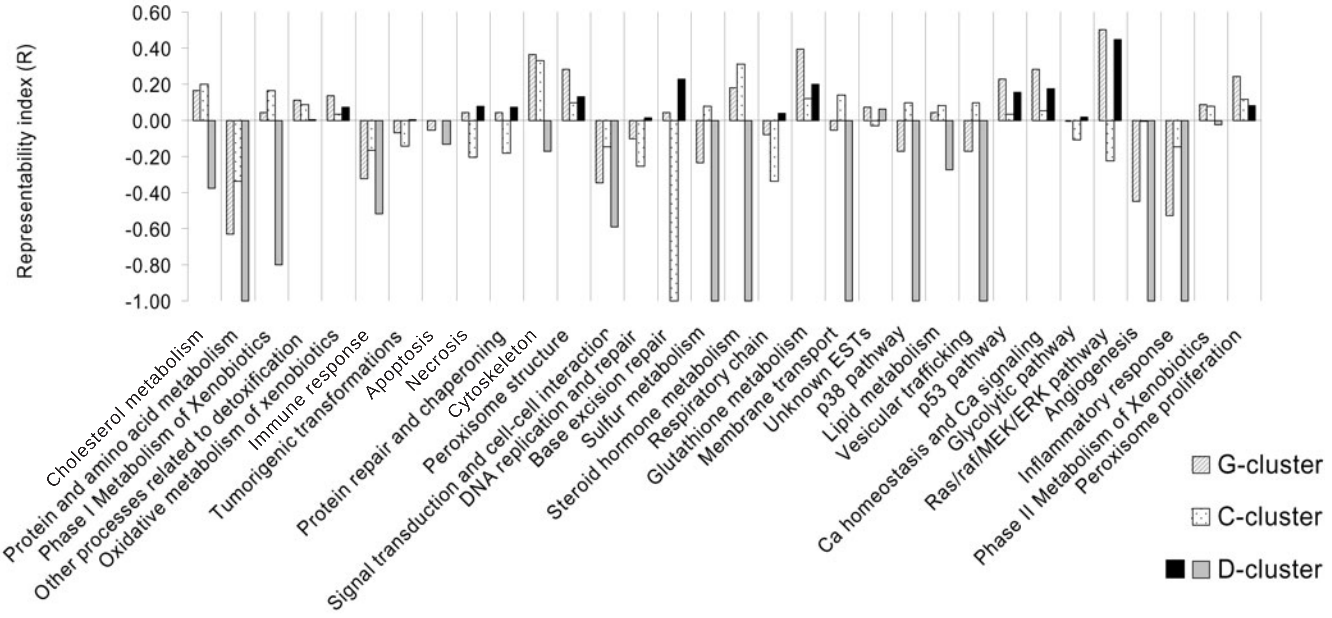

Fig. 1. Representation indices for Pathways class entities for the response of rat hepatocytes to hydrazine treatment. The G-cluster is composed of all genes that have statistically significant changes in their expression as a result of hydrazine treatments in comparison to Time-Matched Controls. The C-cluster is a subset of the G-cluster and includes genes with similar dynamic patterns for all three experimental groups. The D-cluster includes genes of the G-cluster with at least one different dynamic pattern among the three experimental groups. All three clusters were analyzed by queries to the database to categorize genes into entities of the Pathways class. Entities in the D-cluster with positive R_{ij}-values are colored in black and identify those cellular processes that may be more specific for the cellular response of rat hepatocytes to hydrazine treatment. The following entities of the Pathways class are not represented in the G-cluster ($R_{ij} = -1$) and, therefore, are not affected by hydrazine treatment: ribosome structure proteasome structure; iron homeostasis; urea cycle; nucleotide metabolism; growth hormone metabolism; ubiquitin pathway; catexolamine metabolism; trafficking between nucleus and cytoplasm; cell junctions; gonadogenesis; thyroid hormone metabolism; melatonin synthesis; hem synthesis; formation of the extracellular matrix; JNK/SAPK pathway; ATP-sensitive potassium channel; structural component of extracellular matrix.

expression with time, suggesting that simply culturing hepatocytes, even without hydrazine, depresses regular metabolic and detoxification processes in rat hepatocytes. Trends in the expression of these genes show however that treatment with hydrazine does tend to depress the expression somewhat beyond that observed in the Time-Matched Control.

The most interesting group of pathways and functions includes entities with representation indexes in D-cluster greater than in G- and C-clusters. These entities indicate those cellular processes and functions that are specific for the stress response of hepatocytes to hydrazine. Analysis of genes in these entities and the time course of their expression reveal that most of them are positively induced after treatment of hydrazine with the concentration 50 mM (Group-50). The concentration 75 mM (Group-75) mainly depresses the activation of these genes with time. From the identified cellular processes, involved genes and trends in their expression (Table 4), we summarized the stress response induced by hydrazine in hepatocytes.

(1) *Induction of oxidative stress accompanied by disruption of glutathione metabolism and peroxisome function:* This effect is revealed by (i) up-regulation of heat shock proteins, which correlates with oxidative stress (Arrigo, 1998) and (ii) a down-regulation (especially strong for the exposure concentration 75 mM) of most enzymes involved in glutathione metabolism and peroxisome function that may cause disruption of the cellular antioxidation defense in hepatocytes and oxidative damage of mitochondrial DNA (Rusyn *et al*., 2000; Sugiyama *et al*., 2000).

(2) *Perturbations in Ca homeostasis and Ca signaling:* The involvement of Ca regulation is suggested by the down-regulation of calmodulin gene and up-regulation of several other genes that encode calcium-binding proteins, including the molecular chaperons calreticulin and calnexin. It was shown that alterations in Ca binding and signalling also might relate to mitochondrial damage (Xu *et al*., 2001).

(3) *DNA damage with oxidative base modifications and DNA strand breaks:* An effect on DNA integrity is indicated by up-regulation of several DNA repair enzymes including DNA polymerase beta,

Table 4. Gene products identified in over-represented entities by the Functions and Pathways class analysis of the D-cluster and trends in their expression revealed by *post hoc* analysis of gene expression data

Entities of the classes Functions and Pathways	Gene products and number of their probes in the D-cluster (if it is more than 1)	Changes in the expression of the indicated genes in comparison to pretreatment control
	Functions class analysis	
Transcriptional factors	Nuclear factor kappa B (NF-κB), p53, GADD45, c-myc, junB, c-jun, fos-related antigen 2, mdm2	Up-regulation except junB and erratic changes for NF-κB
Antioxidant	Glutathione reductase (2) and peroxidase (2)	Down-regulation, especially strong for Group-75
Chaperons	HSP70(3); HSP60(2); HSP40; DnaJ (glucose-regulated proteins); calreticulin; calnexin	Up-regulation for Group-50 and erratic changes for Group-75
Caspase	Nedd2/Ich-1	Up-regulation
Mitogen-activated protein kinases	c-raf; extracellular signal-regulated kinases ERK (2)	Up-regulation for c-raf and erratic for one probe of ERK
Other enzymes involved in detoxification	Epoxide hydrolase; cyclophilin B; rapamycin; bleomycin hydrolase (2); catechol-*O*-methyltransferase	Erratic changes
Apoptosis inhibitor	bcl-2 (3)	Up-regulation with positive dynamic pattern
	Pathways class analysis	
Respiratory chain	Cytochrome oxidase (2), NADH ubiquinone oxidoreductase, cytochrome c, Rieske iron–sulfur protein (2)	Down-regulation except one probe of cytochrome oxidase
Oxidative metabolism of xenobiotics	Cytochrome b5; NAD(P)H-menadione oxidoreductase; epoxide hydrolase; catechol-*O*-methyltransferase; enzymes included in the entities Respiratory chain and Glutathione metabolism.	Erratic changes
Protein repair and chaperoning	HSP70(3); HSP60(2); HSP40; DnaJ; calreticulin; calnexin; calmodulin	Up-regulation except calmodulin
DNA replication and repair	DNA polymerase beta; AP endonuclease; GADD45; glyceraldehyde-3-phosphate-dehydrogenase; poly(ADP-ribose) polymerase	Up-regulation for Group-50 and erratic for Group-75
Base excision repair	DNA polymerase beta; AP endonuclease	Up-regulation, erratic for Group-75
Peroxisome structure and proliferation	Peroxisome assembly factor-1 (2), peroxisome assembly factor-3; multidrug-resistance transporter P-glycoprotein; vascular endothelial growth factor (VEGF) (2); peroxisome proliferator-induced acyl-CoA thioesterase; acyl coA oxidase	Down-regulation (especially strong for Group-75) and up-regulation for VEGF
Glutathione metabolism	Glutathione S-transferase (3), peroxidase (2), reductase (2), synthetase; gamma-glutamylcysteine synthetase	Down-regulation except gamma-glutamylcysteine synthetase and 1 probe of Glutathione S-transferase
p53 pathway	p21 protein; bcl-2 (3); mdm2; p53 (2)	Up-regulation for Group-50 and erratic for Group-75
Ca homeostasis and Ca signaling	Cyclophilin B; calreticulin and calnexin; calmodulin; rapamycin and FKBP12 target-1 protein	Up-regulation except calmodulin and FKBP12 target-1 protein
Ras/raf/MEK/ERK pathway	c-raf protooncogene; AP-1 transcriptional factors (5); fos-related antigen 2 (3); ERK (2); calmodulin	Up-regulation except calmodulin and one probe of ERK
Necrosis	poly(ADP-ribose) polymerase, calreticulin; calnexin	Up-regulation for Group-50

AP endonuclease and poly(ADP-ribose) polymerase. The first two genes are involved in the Base Excision Repair pathway (Bohr and Dianov, 1999), while the third has been related to DNA strand breaks and necrotic cell death (Masson *et al*., 1995; Ha and Snyder, 1999).

(4) *Alterations in cell cycle control:* Following hydrazine treatment, an up-regulation of nuclear oncoprotein p53, its inhibitor mdm2 and cyclin-dependent kinase inhibitor gene p21 is observed. Crosstalk between these genes is a crucial component of the *p53 pathway* (Hellin *et al*., 2000).

(5) *Induction of the Ras/raf/MEK/ERK pathway:* The involvement of this pathway is revealed by the up-regulation of several proteins involved in this pathway and by down-regulation of calmodulin, which modulates this signaling by its inhibitory effect on Ras activation (Shields *et al*., 2000).

(6) *Changes in apoptosis regulation:* The impact of hydrazine treatment on the regulation of the apoptosis pathway is indicated by upregulation of AP-1 transcriptional factors including c-myc and antiapoptotic Bcl-2 proteins (Goldsworthy *et al*., 1994; Yamamoto *et al*., 1998). These proteins are downregulated in the untreated experimental group. Interestingly, Bax, another well-known member of Bcl-2 protein family, but with pro-apoptotic activity, is stably upregulated and its expression has a very similar positive dynamic with time in all experimental groups.

The cellular processes identified above using the database agree with previous studies on the effect of hydrazine. Hydrazine toxicity in hepatocytes has been associated with the generation of free radicals and the formation of megamitochondria (Matsuhashi *et al*., 1997), carcinogenicity and mutagenicity (Sax, 1990), deactivation of enzymes involved in the antioxidant defense (Kutsenko *et al*., 2001), oxidative stress and apoptosis (Hussain and Frazier, 2002, 2003). In addition, the presented approach allowed us to reveal several other processes involved in the toxic response, such as alterations in cell cycle control, perturbations in Ca homeostasis and Ca signaling, and changes in the regulation of apoptosis, that have not previously been associated with hydrazine and hepatocytes.

DISCUSSION

Advantages and limitations

The tailored database presented here is an example of a tool destined for a local use. If compared with the GO approach, this database has a relational structure and vocabulary tailored for a specific task. This vocabulary composed of eight entity classes allows the automatic elucidation of biological similarity within clusters of genes identified by mathematical analysis of gene expression data. It also allows the numerical estimate of the representation of different cellular functions, pathways and localizations of gene products within these clusters. As demonstrated by the example, the database also allows one to obtain biological clusters of genes according to the cellular functions of the encoded proteins, their localizations or the involvement in pathways. Furthermore, the database serves as an immediate source of focused biological information about a gene of interest, because it incorporates not only structurally organized information about cellular functions, pathways and localizations, but also provides text-based descriptions of all gene products in relation to a specific task, in this case a stress response of rat liver cells to toxicants. Users may easily supplement this database by incorporating new classes and new entities in the existing classes or new information in the text entries.

The absence of complex child–parent relationships is both strength and a limitation of the presented database model. The limitation is that the single-level classes incorporate entities at different levels of specification. For example, some entities in class Pathways represent highly specific pathways, while other entities are more general. While this approach may seem to lack rigor, in fact it reflects the challenge any categorization scheme faces when dealing with intracellular biological processes. The reality is that these processes are complex, interconnected networks of events, and forcing them into a well-defined categorical tree is impossible.

Another concern is the element of subjectivism always present in the categorization and annotation of genes. The selection of class entities and their annotation is highly dependent on the availability of adequate research data and on the completeness of the collection process. Even when there is a large accumulation of data for a given gene, logical inferences made for a particular gene category or annotation by the database curator depends on the individual. This problem can only be minimized by careful attention to details within the context of well-defined quality assurance criteria, and by verification and validation of categories and annotations.

The current version of the database is designed to facilitate the biological interpretations of gene expression data obtained in toxicological experiments using a specific cell culture system (primary cultures of rat hepatocytes) and a specific gene array (the Affymetrix Rat Toxicology U34 GeneChip). This specific focus has advantages and disadvantages. The advantage is that the interpretation is highly relevant to understanding hepatotoxicity without extraneous information about other systems that may or may not be relevant. The disadvantage is that the annotations of gene products are not complete. They were dictated by the constitution of the array and objectives of the study. Therefore, the database may not be as useful for experiments with another cell type or another expression microarray platform.

Our approach thus focuses on creating an easy-to-use tailored database that provides an overview of key function, pathway and localization entities for a given cluster of genes and an estimate of their representation in the cluster. As was demonstrated with the example, the general picture of hepatocyte response to hydrazine obtained by queries to the database provides insights into the mechanism of the toxic effect and reveals directions of changes in cellular processes and functions. The characterization of the mode of action of a toxicant at the level of involved pathway and functions by means of their representation indexes will provide a view distinct from that of toxicity fingerprints based on individual genes.

Perspectives on further development

The primary issue with further development is the tradeoff between the tailored approach, which admittedly is labor intensive, and the generalized, but efficient and systematic, approach implemented by GO. One possible outcome of this study is to highlight the type of realistic, problem-centered investigation that one would like to pursue with a gene array database. The general nature of GO categories was found to be less useful for data interpretation than the specific toxicological focus of our database. The ideal solution would be to have the best of both worlds, efficient and comprehensive like GO, but with the ability to present biological data tailored to a field such as toxicology. While a challenging task, we feel that future databases could move in this direction. One could conceivably add metabolic pathway information specific to disease, toxicology and species to GO, and then with appropriate filters obtain results very similar to what a tailored database provides without the large and inefficient manpower investment needed for each laboratory to build their own database.

In addition, the database currently does not include information on the interactions between gene products integrated in a pathway. However, the categorization of genes in the class Pathways significantly facilitates the incorporation of this information if necessary. A major limitation along these lines is the limited number of genes present on the Rat Tox U34 GeneChip, and therefore, it is impossible to reconstruct many essential interactions. Some important stress related pathways are represented by only several genes in the array. The use of whole genome array should eliminate this limitation.

Many other directions for the development of the database are possible, including the incorporation of information on specific toxicants, effects on different species, manifestations of toxicity at the level of the whole organism and disease states. Addition of these new classes will promote the application of gene array analysis in: (i) understanding the mode of action of chemicals and the identification of 'fingerprints' of their toxicity; (ii) the simulation of biological mechanisms of toxicity and their manifestations on a qualitative basis; (iii) quantitative assessment of dose effect; (iv) the prediction of toxicity and toxic damages of chemicals at the whole organism level; (v) the search for environmental susceptibility genes and (vii) the development of new 'toxchips'.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance of Dr Victor Chan (ManTech Environmental Technologies, Inc.) and Dr Charles Wang (Cedars Sinai Medical Center) with several phases of this project. This study was supported by grants from the Dayton Area Graduate Studies Institute and the Air Force Office of Scientific Research.

REFERENCES

Arrigo,A.P. (1998) Small stress proteins: chaperones that act as regulators of intracellular redox state and programmed cell death. *Biol. Chem.*, **379**, 19–26.

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. *Nat. Genet.*, **25**, 25–29.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.

Bohr,V.A. and Dianov,G.L. (1999) Oxidative DNA damage processing in nuclear and mitochondrial DNA. *Biochimie*, **81**, 155–160.

Bono,H., Kasukawa,T., Hayashizaki,Y. and Okazaki,Y. (2002) READ: RIKEN Expression Array Database. *Nucleic Acids Res.*, **30**, 211–213.

Chang,B., Somogyi,R. and Fuhrman,S. (2001) Evidence for shared genetic programs from cluster analysis of hippocampal gene expression dynamics in development and response to injury. *Restor. Neurol. Neurosci.*, **18**, 115–125.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,L. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al*. (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

Goldsworthy,T.L., Goldsworthy,S.M., Sprankle,C.S. and Butterworth,B.E. (1994) Expression of myc, fos and Ha-ras associated with chemically induced cell proliferation in the rat liver. *Cell Prolif.*, **27**, 269–278.

Gough,N.R. and Ray,L.B. (2002) Mapping cellular signaling. *Sci. STKE*, **135**, EG 8.

Guillouzo,A. (1998) Liver cell models *in vitro* toxicology. *Environ. Health Perspect.*, **2**, 511–532.

Ha,H.C. and Snyder,S.H. (1999) Poly(ADP-ribose) polymerase is a mediator of necrotic cell death by ATP depletion. *Proc. Natl Acad. Sci., USA*, **96**, 13978–13982.

Hellin,A.C., Bentires-Alj,M., Verlaet,M., Benoit,V., Gielen,J., Bours,V. and Merville,M.P. (2000) Roles of nuclear factor-kappaB, p53, and p21/WAF1 in daunomycin-induced cell cycle arrest and apoptosis. *J. Pharmacol. Exp. Ther.*, **295**, 870–878.

Hussain,S.M. and Frazier,J.M. (2002) Cellular toxicity of hydrazine in primary rat hepatocytes. *Toxicol. Sci.*, **69**, 424–432.

Hussain,S.M. and Frazier,J.M. (2003) Involvement of apoptosis in hydrazine induced toxicity in rat primary hepatocytes. *Toxicol*. In Vitro, **17**, 343–355.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.

Keller,W.C. (1988) Toxicity assessment of hydrazine fuels. *Aviat. Space Environ. Med.*, **59**, 100–106.

Kerr,M.K. and Churchill,G.A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, **77**, 123–128.

Kutsenko,S.A., Karpishchenko,A.I., Basharin,V.A. and Glushkov,S.I. (2001) The glutathione-dependent antioxidant system in rats during acute intoxication by hydrazine. *Aviakosm Ekolog. Med.*, **35**, 68–73 (in Russian).

Li,H., Luan,Y., Hong,F. and Li,Y. (2002) Statistical methods for analysis of time course gene expression data. *Front. Biosci.*, **7**, 90–98.

Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.

Masson,M., Rolli,V., Dantzer,F., Trucco,C., Schreiber,V., Fribourg,S., Molinete,M., Ruf,A., Miranda,E.A., Niedergang,C. *et al*. (1995) Poly(ADP-ribose) polymerase: structure–function relationship. *Biochimie*, **77**, 456–461.

Masys,D.R., Welsh,J.B., Lynn Fink,J., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17, 319–326.

Matsuhashi,T., Liu, X., Karbowski,M., Wozniak,M., Antosiewicz,J. and Wakabayashi,T. (1997) Role of free radicals in the mechanism of the hydrazine-induced formation of egamitochondria. *Free Radic. Biol. Med.*, **23**, 285–293.

Raychaudhuri,S., Chang,J.T., Sutphin,P.D. and Altman,R.B. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, **12**, 203–214.

Rusyn,I., Denissenko,M.F., Wong,V.A., Butterworth,B.E., Cunningham,M.L., Upton,P.B., Thurman,R.G. and Swenberg,J.A. (2000) Expression of base excision repair enzymes in rat and mouse liver is induced by peroxisome proliferators and is dependent upon carcinogenic potency. *Carcinogenesis*, **21**, 2141–2145.

Sax,N.I. (1990) Chemical review: hydrazine. *Dangerous Prop. Ind. Mater. Rep*., **10**, 21–58.

Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 47–49.

Schoof,H., Zaccaria,P., Gundlach,H., Lemcke,K., Rudd,S., Kolesov,G., Arnold,R., Mewes,H.W. and Mayer,K.F. (2002) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.*, **30**, 91–93.

Shields,J.M., Pruitt,K., McFall,A., Shaub,A. and Der,C.J. (2000) Understanding Ras: ‘it ain’t over ’til it’s over’. *Trends Cell Biol.*, **10**, 147–154.

Sugiyama,K., Kawamura,A., Izawa,S. and Inoue,Y. (2000) Role of glutathione in heat-shock-induced cell death of *Saccharomyces cerevisiae*. *Biochem. J.*, **352**, 71–78.

Takai-Igarashi,T. and Kaminuma,T. (1999) A pathway finding system for the cell signaling networks database. *In Silico Biol.*, **1**, 129–146.

The FlyBase Consortium (2002) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **30**, 106–108.

Waters,M., Boorman,G., Bushel,P., Cunningham,M., Irwin,R., Merrick,A., Olden,K., Paules,R., Selkirk,J., Stasiewicz,S., Weis,B., Houten,B.V., Walker,N. and Tennant,R. (2003) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) Knowledge Base. *Environ. Health Perspect.*, **111**, 811–824.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S. and Eisenberg,D. (2002) DIP, The Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *NAR*, **30**, 303–305.

Xie,H., Wasserman,A., Levine,Z., Novik,A., Grebinskiy,V., Shoshan,A. and Mintz,L. (2002) Large-scale protein annotation through gene ontology, *Genome Res.*, **12**, 785–794.

Xu,K., Tavernarakis,N. and Driscoll,M. (2001) Necrotic cell death in *C.elegans* requires the function of calreticulin and regulators of Ca(2+) release from the endoplasmic reticulum. *Neuron*, **31**, 957–971.

Yamamoto,H., Ohdan,H., Shintaku,S., Miyata,Y., Marubayashi,S., Asahara,T., Ito,H. and Dohi,K. (1998) Expression of Bcl-2/Bax mRNA in grafted liver during acute rejection after rat liver transplantation. *Transplant Proc.*, **30**, 2950–2951.