BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 3080–3086
doi:10.1093/bioinformatics/btt531

Databases and ontologies

Advance Access publication September 17, 2013

Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform

Johan Nyström-Persson[1,*], Yoshinobu Igarashi[1,*], Maori Ito[1], Mizuki Morita[1,2], Noriyuki Nakatsu[1], Hiroshi Yamada[1] and Kenji Mizuguchi[1,*]

[1]National Institute of Biomedical Innovation, 7-6-8 Saito-Asagi, Ibaraki City, Osaka 567-0085, Japan and [2]Centre for Knowledge Structuring, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Associate Editor: Janet Kelso

ABSTRACT

Motivation: In early stage drug development, it is desirable to assess the toxicity of compounds as quickly as possible. Biomarker genes can help predict whether a candidate drug will adversely affect a given individual, but they are often difficult to discover. In addition, the mechanism of toxicity of many drugs and common compounds is not yet well understood. The Japanese Toxicogenomics Project provides a large database of systematically collected microarray samples from rats (liver, kidney and primary hepatocytes) and human cells (primary hepatocytes) after exposure to 170 different compounds in different dosages and at different time intervals. However, until now, no intuitive user interface has been publically available, making it time consuming and difficult for individual researchers to explore the data.

Results: We present Toxygates, a user-friendly integrated analysis platform for this database. Toxygates combines a large microarray dataset with the ability to fetch semantic linked data, such as pathways, compound–protein interactions and orthologs, on demand. It can also perform pattern-based compound ranking with respect to the expression values of a set of relevant candidate genes. By using Toxygates, users can freely interrogate the transcriptome's response to particular compounds and conditions, which enables deep exploration of toxicity mechanisms.

Availability and implementation: Toxygates is freely available to the public at http://toxygates.nibio.go.jp.

Contact: johan@nibio.go.jp, kenji@nibio.go.jp or y-igarashi@nibio.go.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 30, 2013; revised on August 22, 2013; accepted on September 9, 2013

1 INTRODUCTION

Toxicogenomics is one of several recent '*omics*' approaches to the safety assessment of chemical compounds in early stage drug development. Gene expression analysis of animals' target organs after drug administration can help assess potential toxicity before phenotypic appearance through the use of biomarker genes (Fielden *et al.*, 2007; Uehara *et al.*, 2008).

The Japanese Toxicogenomics Project (TGP) (Uehara *et al.*, 2010), an initiative that set out to collect a large amount of toxicogenomics data systematically, began in 2002 as a joint goverment–private sector project. This project lasted for 10 years, ending in 2012. Its first outcome was the original *TG-GATEs* (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System), which is not open to the public. TG-GATEs is a data storage and toxicity assessment system for data analysis as part of the project. In early 2011, most of the TGP data, including the gene expression, biochemical, blood and histopathological data, as well as high resolution images of *in vivo* organ samples, were released to the public through the *Open TG-GATEs* (OTG) Web site (http://toxico.nibio.go.jp). However, given the current form of OTG, it is not easy for users to explore these gene expression data on their personal computers, because of the large amount of data as well as its static organization and raw format. For example, to extract correlations between gene expression and variables, such as time or dose level, the user would have to download binary files (in the Affymetrix CEL format), extract human-readable expression values and perform tasks such as normalization before this kind of analysis could begin. To make it more accessible, we now introduce Toxygates, an integrated web-based user-friendly analysis platform.

The TGP data have been produced in accordance with well-planned experimental conditions. There are two main types of data: *in vivo* data and *in vitro* data. The condition pattern of the *in vivo* data, which was collected from *Rattus norvegicus*, is the combination of four time points (3 h, 6 h, 9 h, 24 h), four dose levels (control, low, middle, high) and two organs (liver, kidney). These samples were taken after a single dose of the studied compound was administered. Repeat dose data are also available for *in vivo* experiments. In the case of repeat dose samples, the parameters are identical to the single dose ones, except that the time points are 4 days, 8 days, 15 days and 29 days after the initial administration. The condition pattern of the *in vitro* data, available from *R. Norvegicus* and *Homo Sapiens,* is the combination of three time points (2 h, 8 h, 24 h), four dose levels (control, low, middle, high) and one cell type (primary hepatocytes). These parameters potentially allow investigators to conduct not only classifier analysis, but also time series analysis, dose effect analysis and *vivo–vitro* bridging analysis.

The value of a particular dataset is greater if a larger number of perspectives on it are available. It is simply possible to answer more questions when datasets can be related to each other.

*To whom correspondence should be addressed.

Clearly, it is desirable to integrate OTG with other datasets to the greatest possible extent. In Toxygates, data such as pathways, proteins and compound targets from a variety of databases are integrated and displayed alongside the microarray data, and the basic architecture is designed to allow for easy integration of additional data sources in the future. This is achieved in part by making use of linked semantic data in the resource description format (RDF) (Berners-Lee *et al.*, 2001) (http://www.w3.org/rdf/). RDF makes it easy to combine and evolve datasets in an *ad hoc* fashion while still obtaining well-defined results. Thus, Toxygates is a rich analysis platform that allows for easy investigation of the OTG data, offering several different kinds of analysis functions and data perspectives as well as extensible data integration.

2 RESULTS

Toxygates allows users to explore the available data interactively, optionally downloading them at any stage of processing. Analysis begins by choosing a *dataset*. A dataset is a group of related DNA microarrays (targeting messenger RNA) with a common species, dose type (single or repeat) and organ (if *in vivo*) or *in vitro* cell type. For example, one dataset would be rat/*in vivo*/liver/single dose (RVLS). Table 1 gives an overview of the six datasets in Toxygates.

Data are viewed from the perspective of *sample groups*; we define a sample group as a collection of samples. Typically, each experiment, specified by a combination of the compound, dose level and measurement time point, was done in biological triplicate (for the *in vivo* data), and thus such a combination includes three samples (microarrays). For example, a sample group can be defined as consisting of **methapyrilene, high dose, 24 h** and **acetaminophen, middle dose, 24 h**. In the RVLS dataset, this group would contain six samples: three from each compound, dose and time combination.

Such groups can easily be contrasted with each other, and genes can be analyzed for differential expression. Microarray expression values are stored in a prenormalized format (see Methods for the normalization details) and can easily be integrated with annotations from external datasets, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa *et al.*, 2012), Gene Ontology terms (Gene Ontology Consortium, 2000) and ChEMBL (Gaulton *et al.*, 2012) or DrugBank (Knox *et al.*, 2011) compounds. Such annotations are obtained dynamically from linked data (Resource Description Format, RDF) endpoints. These tools economize the researchers' time and enable them to proceed to more detailed analysis more quickly.

2.1 User interface

Toxygates is structured around a loose workflow consisting of six different screens. For the most part, it is assumed that the user wants to proceed from known compounds and experimental conditions to extraction of interesting genes and proteins (corresponding to microarray probes). However, by using the *compound ranking* feature, it is also possible to proceed from genes or proteins to compounds. By alternating between these two analysis types, users can iteratively refine their initial hypothesis and gather more evidence.

Table 1. Microarray datasets in toxygates

Species	Cell type	Dose type	Samples	Compounds
Human	*In vitro*	Single	2593	157
Rat	*In vitro*	Single	3370	145
Rat	*In vivo*, liver	Single	6765	143
Rat	*In vivo*, liver	Repeat	7378	158
Rat	*In vivo*, kidney	Single	1952	41
Rat	*In vivo*, kidney	Repeat	1953	41
Total			24 011	170

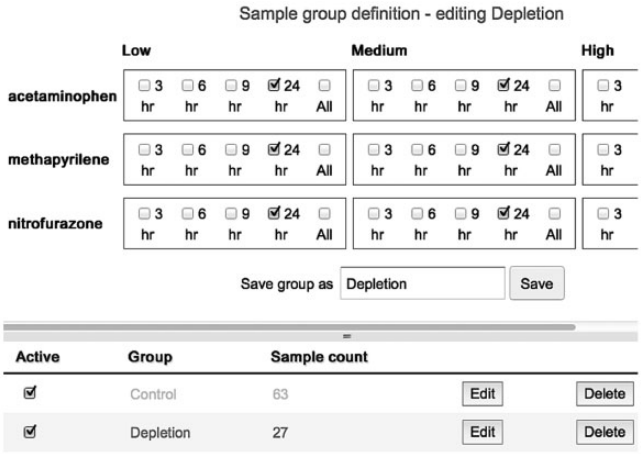

Fig. 1. The sample group definition screen

The six main screens are the following:

- **Dataset selection.** On this screen, the user selects a basic dataset to examine, such as RVLS (see above). In general, data analysis in Toxygates is performed within a single dataset. Because of the lack of a standard procedure for comparing cells from different species, or *in vivo* data with *in vitro* data or repeated with single doses, we do not perform such comparisons automatically. However, by using the tools available in Toxygates, users can easily carry out such comparisons manually.
- **Sample group definitions.** On this screen, the user defines sample groups by selecting first compounds, and then sample times and dose levels that they are interested in (Fig. 1). On this screen, compound ranking can also be performed (to be discussed below).
- **Probe selection.** On this screen, the user can optionally choose the microarray probes they would like to view in the next step. If the user does not want to select a subset of the available probes, they may simply skip this step to view all probes (about 31 000 for rat and 50 000 for human). Probes can be selected based on annotations such as Gene Ontology terms, KEGG pathways, ChEMBL and

DrugBank compound targets and so on. In the latter three cases, data are obtained from remote RDF endpoints (see Methods for implementation details). Toxygates was designed to let us add new probe information sources easily in the future, to further assist probe filtering and thus toxicological hypothesis testing.

- **View data.** This screen is where the main data inspection and analysis is performed. The user can view expression values for the groups that they have defined by viewing pages of 25–100 probes (the amount is user configurable) at a time. Welch's *t*-test, Mann–Whitney *U*-test and the fold-change difference columns can be added to identify the most significant probes for discriminating between any two different groups. It is possible to display or hide a number of *dynamic columns*, which fetch associations such as KEGG pathways and EggNOG (Powell *et al.*, 2012) orthologous proteins from remote RDF datasets, displaying them alongside the related expression values. Of the 170 compounds in Toxygates, some are known to interact directly with rat or human proteins, as annotated in ChEMBL and DrugBank. These compounds are also displayed as dynamic columns alongside genes that produce the target proteins. As with probe annotation sources, it is easy for us to add additional dynamic columns over time as the need arises. In addition to these features, the user can view charts to inspect the behavior of a certain probe visually, as well as export the data (under the Action menu) to comma-separated value (CSV) files or to the TargetMine system (Chen *et al.*, 2011) for further analysis, such as biological enrichment, a type of analysis that Toxygates does not currently perform.
- **Pathologies.** This screen displays the pathologies that were found in those samples that are part of the current user-defined groups. Pathologies are described using a controlled vocabulary that is specific to OTG (http://toxico.nibio.go.jp/open-tggates/doc/pathology_parameter.pdf).
- **Sample details.** This screen displays all metadata associated with those samples that are part of the current user-defined groups, such as blood composition and biological data (for *in vivo* samples) and the experimental conditions.

2.2 Example: glutathione depletion analysis

As an example use case, we next show how to apply Toxygates to the study of glutathione depletion. The liver and the kidney are major detoxification organs. In the liver, glutathione, which scavenges reactive oxygen species, is one of the major detoxification players. It conjugates target toxic compounds using its thiol group and exports the conjugated compounds into bile ducts. Thus far, two toxicological mechanisms have been reported for drug-induced glutathione depletion. One is inhibition of glutathione synthesis, and the other is excessive usage of glutathione. An overdose of glutathione depleting compounds, such as acetaminophen, may lead to hepatotoxicity through depletion of glutathione (Gao *et al.*, 2010). Here, by using Toxygates, we explore the differentially expressed genes that are affected by glutathione depleting compounds, which are known in advance. Toxygates helps us confirm the hypothesis as well as gather additional evidence for it and discover related compounds.

As glutathione depleting compounds, we selected acetaminophen, methapyrilene and nitrofurazone. We defined a sample group for this positive set using the data points at 24 h of all three dose levels. The 24 h time point was chosen, as the toxicological effects of each compound are visible more clearly here compared with 3 h, 6 h and 9 h. Although Toxygates provides six different datasets (shown above), in this example we focus only on rat/*in vivo*/liver/single data for space reasons. As non-glutathione affected compounds, we selected erythromycin, gentamicin, glibenclamide, hexachlorobenzene, isoniazid, penicillamine and perhexilline and defined a sample group for this negative set using the data points at 24 h of all three dose levels (Fig. 1). After defining these two sample groups, we may focus only on the probes corresponding to relevant genes, to obtain a more helpful perspective. In this example, we selected the 42 probes belonging to the *glutathione metabolism* pathway. After making this selection, we can display the log-2 ratio fold change of the individual probes. By clicking 'Add T-test', we can perform Welch's *t*-test to compare the mean of the two sample groups and add a column to show the *P*-values calculated. The top four ranked probes are Mgst2 (microsomal glutathione S-transferase 2), G6pd (glucose-6-phosphate dehydrogenase), Gsr (glutathione reductase) and Gclc (glutamate–cysteine ligase). Gclc is known to accelerate glutathione synthesis, and Gsr and G6pd are involved in the conversion of glutathione from the oxidized form to the reduced one. Mgst2 is glutathione-S-transferase, which is the main enzyme for detoxification of toxic compounds by the conjugation reaction.

By clicking the icon on the very left side, we can view the temporal and dose changes of the values graphically (Fig. 2).

2.3 Compound ranking

For the most part, the features mentioned previously help the user discover interesting genes, proteins or probes or annotations such as pathways, for some pre-selected compounds and experimental conditions. However, by using the compound ranking feature, which is available on the sample group definition screen, it is also possible to go from genes or probes to compounds. Compound ranking assigns a match score to each compound and sorts the compound list by descending score. Ranking is done on the basis of match rules, each of which consists of a match type and a gene, the time series of which will be the basis of the matching. The end result of the ranking will thus highlight compounds where a single dose level fits the specified behavior of the specified genes as closely as possible (Fig. 3). Precise details of the ranking methods are given in the Supplementary Material.

As an example, we apply compound ranking to the glutathione depletion data that were inspected in the previous section. The two probes in the *glutathione metabolism* pathway with the smallest *P*-values were G6pd and Mgst2. We define two ranking rules based on these genes. By inspecting the charts on the data screen, we find that glutathione depleting compounds tend to upregulate these genes. Thus, initially we select *total upregulation* as the match type for both. With these rules, the top nine matching compounds include the original three that we used in our group definition (Table 2). The scores simply reflect the sum of all positive log-2 folds. We also attempt a second ranking by

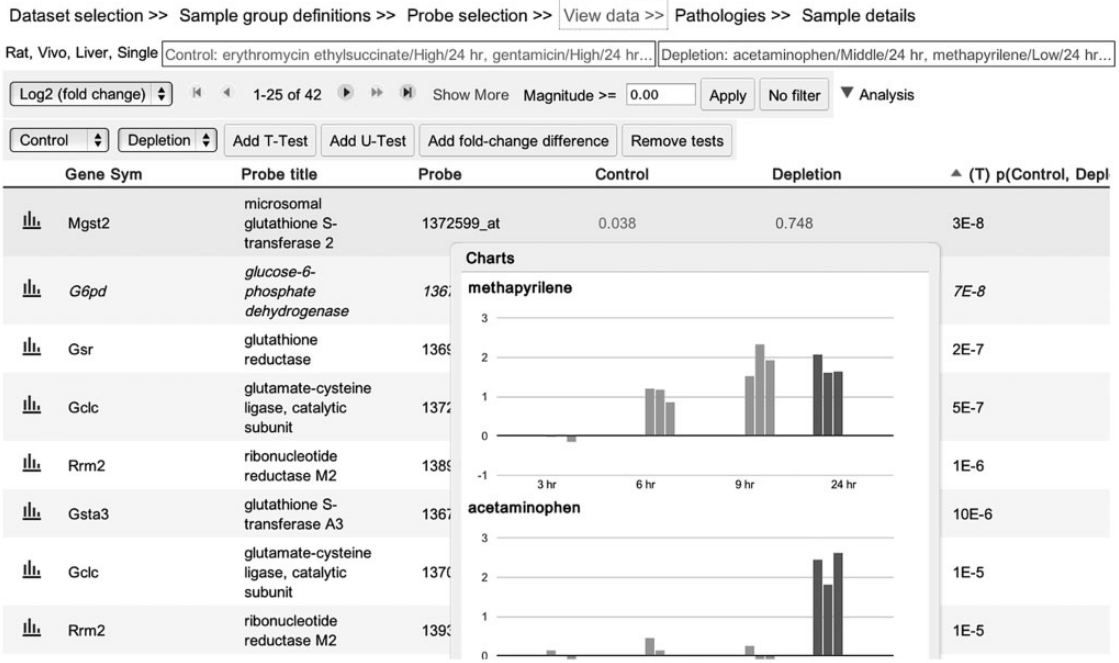

Fig. 2. By clicking the chart icon on the left of each data row, a configurable popup window with dose and time series charts corresponding to the chosen probe can be displayed

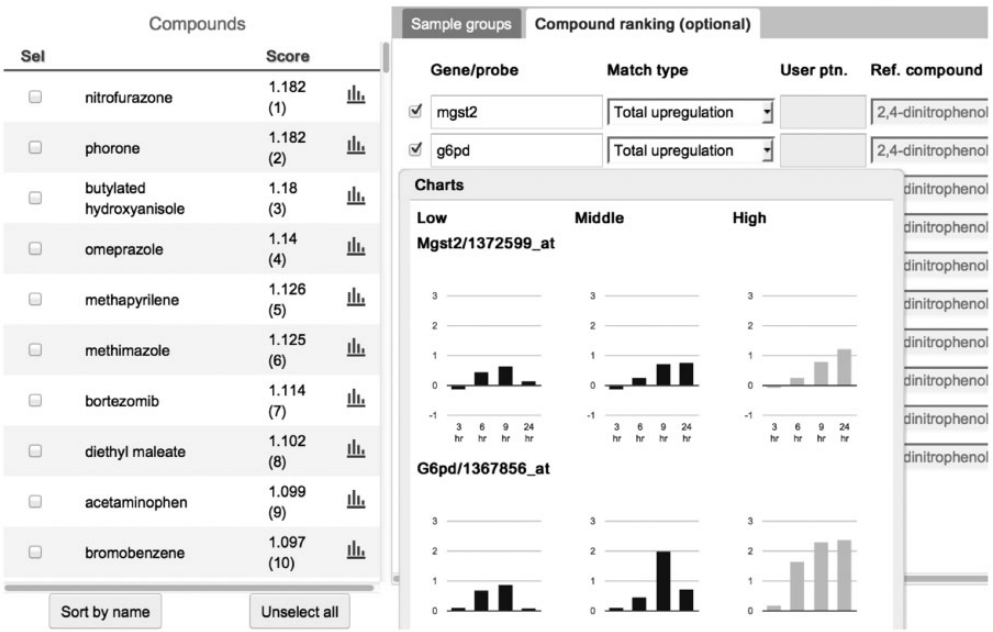

Fig. 3. Compounds have been ranked by their behavior with respect to the Mgst2 and G6pd genes

using the *reference compound* rule, which attempts to find compounds that behave similarly to a given reference. By attempting rankings based on each of our original three compounds, we find that using nitrofurazone and high dose as the reference ranks all three highest. In this case, the scores (Table 3) are based on Pearson correlation and, therefore, cannot be compared numerically with the scores from upregulation.

Only the highest scoring 14 of all 158 compounds for RVLS have been shown. Out of these 14, 9 are shared between the two ranking methods. H/M/L in the parentheses indicates the best matching dose. We also indicate whether the compound is known to conjugate with glutathione (GSH) and deplete GSH, respectively. Given that we are aiming to find compounds that deplete GSH by conjugation, the top ranked compounds can

Table 2. Ranking results when looking for upregulation of the G6pd and Mgst2 genes

Rank	Compound (dose level)	Score	GSH conjugation	GSH depletion
1	*Nitrofurazone* (H)	1.182	Y	Y
2	Phorone (H)	1.182	Y	Y
3	Butylated hydroxyanisole (M)	1.18	N	N
4	Omeprazole (H)	1.14	Y	Y
5	*Methapyrilene* (H)	1.126	Y	Y
6	Methimazole (H)	1.125	?	Y
7	Bortezomib (M)	1.114	?	Y
8	Diethyl maleate (H)	1.102	Y	Y
9	*Acetaminophen* (M)	1.099	Y	Y
10	Bromobenzene (M)	1.097	Y	Y
11	Coumarin (H)	1.096	Y	Y
12	Cephalotin (H)	1.092	?	?
13	Nitrosodiethylamine (H)	1.071	?	Y
14	Phenacetin (M)	1.064	Y	Y

Note: We also indicate reported GSH conjugation and depletion. The letter 'Y' signifies that the compound is known to conjugate with or deplete GSH, respectively. The letter 'N' signifies that it is known not to do this.

Table 3. Ranking results when using nitrofurazone as a reference

Rank	Compound (dose level)	Score	GSH conjugation	GSH depletion
1	*Nitrofurazone* (H)	4	Y	Y
2	Butylated hydroxyanisole (L)	3.939	N	N
3	*Methapyrilene* (H)	3.932	Y	Y
4	Methimazole (H)	3.89	?	Y
5	Carbon tetrachloride (H)	3.881	?	Y
6	Bromobenzene (M)	3.746	Y	Y
7	Omeprazole (H)	3.503	Y	Y
8	Propylthiouracil (H)	3.433	N	N
9	Simvastatin (M)	3.383	?	?
10	Coumarin (H)	3.378	Y	Y
11	Bortezomib (M)	3.282	?	Y
12	Rotenone (M)	3.252	N	Y
13	Thioacetamide (L)	3.23	?	Y
14	*Acetaminophen* (M)	3.211	Y	Y

Note: We also indicate reported GSH conjugation and depletion. The letter 'Y' signifies that the compound is known to conjugate with or deplete GSH, respectively. The letter 'N' signifies that it is known not to do this.

broadly be classified as true positives, false positives and compounds with unclear or different relations to GSH. True positives are, for example, phorone (van Doorn *et al*., 1978), bromobenzene (Comporti *et al*., 1991; Monks *et al*., 1982), coumarin (Luchini *et al*., 2008), omeprazole (Weidolf *et al*., 1992), diethyl maleate (Boyland and Chasseaud, 1967) and phenacetin (Hinson, 1983). False positives are propylthiouracil (Banerjee *et al*., 1998) and butylated hydroxyanisole (Borroz *et al*., 1994). Examples of compounds with unclear or different GSH relation are methimazole (Mizutani *et al*., 1999), carbon tetrachloride (Beddowes *et al*., 2003), bortezomib (Nerini-Molteni *et al*., 2008) and thioacetamide (Lotkova *et al*., 2007). A full classification of the top ranked compounds is given in the Supplementary Material. As would be expected, the top ranking results contain not only compounds whose mechanism is similar but also whose mechanism are different but induce similar expression patterns for focused genes.

As we mentioned previously, glutathione uses its thiol group to the conjugation. The sulfur in the thiol group is provided from L-cysteine amino acids. To gather additional evidence for our hypothesis, an examination of the expression level of other enzymes, which are affected by the amount of L-cysteine, might be valuable. To find other enzymes that use L-cysteine, we surveyed the pathway maps where L-cysteine appears. Furthermore, to simplify the interpretation of data, we extracted only enzymes that irreversibly use L-cysteine as a substrate. Thus we found 'Ppcs' in the coenzyme A biosynthesis pathway. This enzyme acts on L-cysteine directly, and thus would be affected by the amount of L-cysteine. By adding upregulation of Ppcs alongside G6pd and Mgst2 and ranking again, the false-positive compounds like buthionine sulfoximine and ethionine, which do not conjugate glutathione but inhibit Gss enzyme activity, are ranked down from 45th to 56th and 94th to 131st, respectively. Thus, by considering the possible interactions in detail, it is possible to gradually refine the ranking to extract more probable compounds.

2.4 Data architecture

The Toxygates data analysis platform is based on the integration of diverse linked data—obtained from RDF datasets from e.g. SPARQL Protocol and RDF Query Language (SPARQL) endpoints—with a large DNA microarray core dataset. This hybrid—and, we believe, novel—architecture combines the benefits of two different approaches to data integration: data warehousing and linked data (Goble and Stevens, 2008). The former provides high performance for static unchanging data, and the latter provides flexibility for diverse and frequently updated data sources. Because many of the RDF datasets are stored and updated remotely (Supplementary Table S1), updates will transparently feed into Toxygates at the time when they are published.

Many RDF datasets are displayed as *dynamic columns* in our main data screen alongside expression values and probes. Our fundamental design permits us to add additional columns quickly based on data from remote services, as long as SPARQL queries that allow us to look up the information based on gene, protein or probe identifiers can be written.

3 DISCUSSION

3.1 Toxicogenomics databases

A small number of databases that are open to the public and closely related to OTG and Toxygates exist. In general, toxicogenomics data are transcriptomics data of animals that have been treated with drugs or chemical compounds. Thus, transcriptomics information, the kinds and dosages of administered compounds, studied organs and associated information like biochemical parameters other than gene expression data are all essential parts of toxicogenomics datasets. To the best of our knowledge, there exist no other open toxicogenomics databases fully comparable with OTG, if this definition as well as the massive size and the well-documented standard operating procedures of the latter are taken into account.

Comparative Toxicogenomics Database (Davis *et al.*, 2013) is a knowledge base of relations among drugs, genes and diseases that were derived from published literature by curators. However, this database does not contain actual transcriptomics data measurements. Gene Expression Omnibus (Barrett *et al.*, 2011) is the largest repository of gene expression data in the world. This database contains a large amount of gene expression data associated with treatment by bioactive compounds. The number of such experiments is being increased, but the experimental targets and protocols, such as species, strains, administration methods, time points and dose controls are not unified. The lack of such unified experimental conditions can sometimes be a critical obstacle to making efficient use of the data from a toxicological point of view. ArrayExpress (Rustici *et al.*, 2013) is another large expression data repository, maintained by the European Bioinformatics Institute (EBI), with characteristics similar to those of Gene Expression Omnibus. DrugMatrix (Ganter *et al.*, 2005) has *in vivo* rat microarray data for >600 compounds. However, although having just 170 compounds, OTG has a larger and uniform amount of dose levels and time points for each, enabling deep and systematic exploration.

The Connectivity Map (Lamb *et al.*, 2006) is an integrated microarray database where samples are associated with the administration of some compound. Like Toxygates, it permits a form of compound ranking. For pattern matching and rank scoring, the authors have chosen to use the Kolmogorov–Smirnov statistic to gracefully handle situations where several different platforms are present. Because OTG is based on a single platform, Toxygates is able to use a simple Pearson correlation statistic instead.

ToxBank (Kohonen *et al.*, 2013) is a project that aims to provide a comprehensive web-based resource for toxicity research data and protocols, as well as related compound and biomaterial information. It also aims to be a platform for data sharing between investigators. Like Toxygates, it makes use of both linked data and data warehousing approaches. ToxBank has a broader scope than Toxygates, which aims to provide well-integrated tools for microarray data analysis only.

3.2 Data warehousing and linked data

Data integration is a central problem in bioinformatics. However, there are not yet universally accepted solutions for data problems such as varying quality, heterogeneity of identifiers and accessions and unclear semantics. A useful general review of bioinformatics data integration is provided by Goble and Stevens (Goble and Stevens, 2008).

In Toxygates, we combine two approaches with different characteristics. First, *data warehousing* is perhaps the most heavyweight approach but also most efficient if the data perspectives and usage scenarios that are needed can be relied on to be relatively unchanged over time. In this approach, data are scraped from multiple source databases and cleaned to fit together for particular usage scenarios, often by using custom parsers. Data access can be quick and efficient, as all of the data are located in one place. One drawback is that updating the integrated data to a new version when the source databases are updated may require a lot of manual work. A popular data warehousing framework is InterMine (Lyne *et al.*, 2007), which was originally developed for the FlyMine database; it is also used by e.g. TargetMine (Chen *et al.*, 2011).

In recent years, semantic web technologies such as SPARQL, RDF and Web Ontology Language (OWL) have become increasingly popular in bioinformatics. These technologies are examples of *linked semantic data*, the second part of our hybrid approach. Linked semantic data emphasize identifying objects by using Uniform Resource Identifiers (URIs), and encoding resource interrelationships as links to these URIs. Ontologies are used to help make sense of data that has not previously been encountered. We believe that semantic web technologies represent one of the best hopes for bioinformatics data integration so far. On the other hand, the performance of this essentially flexible approach cannot always be as good as that of a static approach like data warehousing. For this reason, Toxygates combines the strengths of both approaches selectively.

4 METHODS

In this section, we discuss the essential data processing methods used and design decisions made for Toxygates. A more extensive discussion is given in the Supplementary Material.

4.1 Microarray data and normalization

Starting from the raw Affymetrix GeneChip data from OTG in CEL format (available from http://toxico.nibio.go.jp), expression and call values were extracted by using the Bioconductor Affy package for R (http://www.bioconductor.org). Two databases were constructed: the absolute value database and the fold database. For absolute values, the **mas5** function was used with the **normalize** flag set to 'true'. The **mas5calls** function was used to extract calls (present/absent/marginal). For fold values, the **mas5** function was used again but with **normalize** set to 'false', and the data were normalized by the median value of each sample. Further details on our treatment of the microarray data are given in the Supplementary Material.

4.2 RDF-ization of OTG

As a preparatory step in the creation of Toxygates and to support integration with semantic linked data, OTG data were converted into RDF format. Included in this conversion were experimental conditions, biochemical, blood and histopathological data, that is, everything except for RNA expression values and image data was converted. Expression values were omitted due to size and efficiency considerations, and image data were omitted due to images having no natural representation as RDF triples (however, in the future, we plan to include links to these images in the data). The result of this conversion was about 2.4 million triples, ~72 triples per sample for 33 566 samples. The number varies, as samples can have different numbers of pathological findings.

Many biological RDF datasets are made available to the public through SPARQL endpoints, which anybody may query on the web. Such endpoints are used extensively by Toxygates to retrieve and display auxiliary information. In this release of Toxygates, as our main focus has been to provide a user-friendly analysis platform, we do not make our sample data available through such an endpoint, although we may do so in the future. Further details on our use of RDF are given in the Supplementary Material.

5 CONCLUSION

We have described Toxygates, a novel user-friendly analysis platform for OTG, a comprehensive toxicogenomics database

developed in Japan during the span of 10 years. OTG is in itself a resource of potentially great value for the understanding of toxicity mechanisms and the identification of biomarker genes, assisting preclinical drug development. Toxygates adds further value by making this data collection accessible to a large number of individual researchers and users, providing unique integrated analysis tools, which include dynamic views of remote data, as well as making data access simpler and easier than previously. Toxygates allows users to query OTG both by going from compounds and experimental conditions to relevant genes, and by going from relevant genes to compounds, although currently we place greater emphasis on the former style of investigation. The current release of Toxygates represents the first public release, and we hope to make several improvements. We aim to curate and expand the collection of remote RDF datasets that have been integrated as probe sources and as dynamic columns. We also aim to integrate other gene expression datasets into Toxygates. Finally, we hope to provide additional valuable perspectives and data analysis methods in response to user feedback.

ACKNOWLEDGEMENTS

The authors are grateful to the maintainers of the ChEMBL SPARQL endpoint at Uppsala University and to the Bio2RDF developers for helping them to resolve issues quickly. They also wish to thank Lokesh Tripathi, Ken Ishii, Cevayir Çoban and the members of the Database Centre for Life Sciences for testing Toxygates and giving them feedback during the development process.

Funding: The Health and Labour Sciences Research Grant 'Adjuvant database Project' of the Japanese Ministry of Health, Labour and Welfare (in part), and a collaborative program between the National Institute of Biomedical Innovation (NIBIO) and Japan's National Bioscience Database Center (NBDC) (in part).

Conflict of Interest: None declared.

REFERENCES

Banerjee,A. *et al.* (1998) Induction of an ATPase inhibitor protein by propylthiouracil and protection against paracetamol (acetaminophen) hepatotoxicity in the rat. *Br. J. Pharmacol.*, **124**, 1041–1047.

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.

Beddowes,E.J. *et al.* (2003) Chloroform, carbon tetrachloride and glutathione depletion induce secondary genotoxicity in liver cells via oxidative stress. *Toxicology*, **187**, 101–115.

Berners-Lee,T. *et al.* (2001) The semantic web. *Scie. Am.*, **284**, 34–43.

Borroz,K.I. *et al.* (1994) Modulation of gamma-glutamylcysteine synthetase large subunit mRNA expression by butylated hydroxyanisole. *Toxicol. Appl. Pharmacol.*, **126**, 150–155.

Boyland,E. and Chasseaud,L.F. (1967) Enzyme-catalysed conjugations of glutathione with unsaturated compounds. *Biochem. J.*, **104**, 95–102.

Chen,Y.A. *et al.* (2011) Targetmine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, **6**, e17844.

Comporti,M. *et al.* (1991) Glutathione depletion: its effects on other antioxidant systems and hepatocellular damage. *Xenobiotica*, **21**, 1067–1076.

Davis,A.P. *et al.* (2013) The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.

Fielden,M.R. *et al.* (2007) A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. *Toxicol. Sci.*, **99**, 90–100.

Ganter,B. *et al.* (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.*, **119**, 219–244.

Gao,W. *et al.* (2010) Mechanism-based biomarker gene sets for glutathione depletion-related hepatotoxicity in rats. *Toxicol. Appl. Pharmacol.*, **247**, 211–221.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids Res.*, **40**, D1100–D1107.

Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genetics*, **25**, 25–29.

Goble,C. and Stevens,R. (2008) State of the nation in data integration for bioinformatics. *J. Biomed. Inform.*, **41**, 687–693.

Hinson,J.A. (1983) Reactive metabolites of phenacetin and acetaminophen: a review. *Environ. Health Perspect.*, **49**, 71–79.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Kohonen,P. *et al.* (2013) The ToxBank Data Warehouse: Supporting the Replacement of In Vivo Repeated Dose Systemic Toxicity Testing. *Mol. Inform.*, **32**, 47–63.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Lotkova,H. *et al.* (2007) *S-adenosylmethionine* exerts a protective effect against thioacetamide-induced injury in primary cultures of rat hepatocytes. *Altern. Lab. Anim.*, **35**, 363–371.

Luchini,A.C. *et al.* (2008) Intestinal anti-inflammatory activity of coumarin and 4-hydroxycoumarin in the trinitrobenzenesulphonic acid model of rat colitis. *Biol. Pharm. Bull.*, **31**, 1343–1350.

Lyne,R. *et al.* (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biol.*, **8**, R129.

Mizutani,T. *et al.* (1999) Metabolism-dependent hepatotoxicity of methimazole in mice depleted of glutathione. *J. Appl. Toxicol.*, **19**, 193–198.

Monks,T.J. *et al.* (1982) Stereoselective formation of bromobenzene glutathione conjugates. *Chem. Biol. Interact.*, **41**, 203–216.

Nerini-Molteni,S. *et al.* (2008) Redox homeostasis modulates the sensitivity of myeloma cells to bortezomib. *Br. J. Haematol.*, **141**, 494–503.

Powell,S. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.

Rustici,G. *et al.* (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.*, **41**, D987–D990.

Uehara,T. *et al.* (2008) A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats. *Toxicology*, **250**, 15–26.

Uehara,T. *et al.* (2010) The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res.*, **54**, 218–227.

van Doorn,R. *et al.* (1978) Synergistic effects of phorone on the hepatotoxicity of bromobenzene and paracetamol in mice. *Toxicology*, **11**, 225–233.

Weidolf,L. *et al.* (1992) A metabolic route of omeprazole involving conjugation with glutathione identified in the rat. *Drug Metab. Dispos.*, **20**, 262–267.