20

Prediction of Hepatotoxicity Based on the Toxicogenomics Database

Tetsuro Urushidani

20.1 Introduction

Today, in the post-genomic era, there have been remarkable advances in the technology of searching for seeds of new drugs. However, the success rate of drug development is nevertheless decreasing not only in Japan but also worldwide. According to the investigation performed by the Japan Pharmaceutical Manufacturers Association (http://www.jpma.or.jp/12english/publications/index.html), 422 653 seeds for new drugs were synthesized or extracted from 1996 to 2000 in Japan, whereas only 63 were approved. Moreover, half of those approved were chemicals created by and purchased from foreign countries. It is obvious that the opportunity to create an original drug is now extremely low.

Drug development in the previous century was usually based on screening by measuring effects of the chemicals in model animals with artificially created diseases, and subsequently it sometimes happened that a quite excellent drug was produced not for humans but for rats. In recent years, however, it has been possible to start the development by targeting disease-related genes whose molecular functions are well elucidated, and indeed, human-type genes are always available. Therefore, it is now easy to select a chemical, which is effective on the human-type molecule for at least the *in vitro* level. Even with this advantage, many candidate drugs have been eliminated largely because of toxicity, which could not be found in pre-clinical tests but appeared in clinical trials (Ismail and Landis, 2003). This not only brings about an increment in the cost for drug development that will be shifted to the medical expense, but also causes a serious ethical problem that the toxicity of the chemical is proven by using humans. From another

Hepatotoxicity: From Genomics to in vitro *and* in vivo *Models* Edited by S. C. Sahu

point of view, the low predictivity of present toxicological tests for clinical toxicity also means that many drug candidates might have been abandoned although they were in fact excellent.

The current toxicological tests are much improved and controlled to assure reliability of the data in comparison to past adverse cases, such as thalidomide. However, reliability of the pre-clinical data does not necessarily lead to clinical safety. When the drug possesses an obvious organ or cellular toxicity as part of its pharmacological effects, its toxicity can be easily predicted from its dose–response relationship and thus controlled. However, many of the practical side effects in the clinical field are scarcely related to the pharmacological target and the origin of the toxicity is usually not the drug itself but its metabolite(s). If a serious adverse effect happened in the clinical dose range at a rate of 1/1000 (an impractically large number), such a drug should be withdrawn from the market at once. However, it is theoretically impossible to detect the effect with such a low incidence using 10 000 rats or more, even though there is no species difference, and this is, of course, still an impractical number for safety tests. Needless to say, the barrier of species difference is quite high and we sometimes encounter a case that there is a difficulty to find out an animal species accurately reproducing a clinically evident toxicological profile. In practical toxicological tests, we inevitably employ the hopeless assumption that any pathophysiological changes in the experimental animals associated with the over-dosage of the test drug can be extrapolated to the corresponding adverse effects with low incidence in humans at low dose. Many conscientious toxicologists feel that such a preclinical safety evaluation is not a certification but an excuse.

In order to 'predict the unpredictable toxicity', the most realistic strategy would be 'toxicogenomics' which enables us to analyze all of the gene expression changes caused by an external stimulus in an organism. Even when an 'unexpected' hazard is elicited by drug administration at a very low incidence, its causal events should have occurred and some of them should be detectable as a subtle change. If all of the events were observed and their relationships were analyzed, a perfect prediction would be realized. Observation of all of the physiological processes had been thought to be practically impossible, but advances in biotechnology have made it possible. Since the most sensitive change among biological processes is that of mRNA, and many of the physiological responses are expected to be associated with changes in the expression level of mRNA, the most promising measure is the amount of mRNA. The development of microarray technology has made it possible to substantially quantify the expression level of all of the genes at once.

Under such a circumstance, the Ministry of Health, Labour and Welfare, National Institute of Health Sciences (NIHS) and the working group of the Japan Pharmaceutical Manufacturers Association planned 'The Toxicogenomics Project (TGP)', a collaborative project of the government and private companies (Urushidani and Nagao, 2005). This is a 5-year project from 2002 to 2007, and after the rearrangements of the organization and mergers of the companies, the members are the NIHS, 15 pharmaceutical companies (Astellas, Chugai, Daiichi, Dainippon-Sumitomo, Eisai, Kissei, Mitsubishi, Mochida, Ono, Otsuka, Sankyo, Sanwa, Shionogi, Takeda and Tanabe) and the National Institute for Biomedical Innovation (NIBIO) as the core institute where the actual work is being performed. About half of the budget is from a grant of the Ministry of Health, Labour and Welfare, and the remaining is from the companies.

20.2 The Features of The Toxicogenomics Project (TGP)

The plan of the project at the beginning was as follows. The main goal is to create a gene expression database of 150 chemicals, mainly medical drugs. The target organ is mainly the liver. The reason was that most of the clinically serious adverse effects have occurred in this organ, and that the composition of the cell type is relatively homogenous in this organ and thus the expected variation based on differences of sampling would be minimal. Subsequently, the liver was considered to be favorable to accumulate the know-how of the toxicogenomics technique. Nephrotoxicity was of course considered to be important and the sampling and pathological examination of the kidney in addition to liver was to be performed in all of the animals. Among the animals showing nephrotoxic phenotypes, up to 20 % of the total compounds were subjected to transcriptome analysis of the kidney.

For the test animal, 6 week- old male Spraque–Dawley rats were employed. At the starting point, the rat genome was not fully analyzed and there was a firm opinion that the mouse should be used because of the enrichment of the genome information. However, all of the classical toxicological tests had been done with rats and there was a vast amount of knowledge accumulated. When we were venturing into a totally new field, we believed that the accumulated knowledge should be used as much as possible. Another advantage was that more data, such as biochemistry and hematology, could be obtained from the rat rather than from the mouse.

The biggest problem at the start of the TGP was the fact that we were behind the large pharmaceutical companies and bioventures in Europe and America who had already started similar projects to create databases of toxicogenomics (Porter *et al.*, 2003; Boverhof and Zacharewski, 2006). Therefore, the following strategy was employed to catch up with the preceding projects. (1) Establishment of quantitativity and reproducibility. Data with excellent quantitativity are acquired using the 'Affymetrix GeneChip' and a new normalization method based on the externally added spike RNAs proportional to the sample DNA contents (Kanno *et al.*, 2006) was employed. (2) Selection of test compounds. The chemicals, mainly medicinal, contain drug candidates withdrawn from development because of their hepato- or nephrotoxicity, that are supplied from the member companies. (3) Bridging between species. In addition to the *in vivo* experiments, exposure to the primary culture of rat liver and to human hepatocyte culture was performed. (4) Enrichment of protocol. Gene expression data possess multi-dose, multi-time points that link to various toxicological measures. These points are discussed in this order below.

20.2.1 Establishment of Quantitativity and Reproducibility

The 'Affymetrix GeneChip', an oligonucleotide array, is known to be superior to other arrays, such as the 'Stanford' type (Wildsmith and Spence, 2003). In the TGP, data with excellent quantitativity are being accumulated under thoroughly optimized and controlled conditions in order to draw their full ability. During the first two years (up to ca. 35 chemicals), the 'Rat Expression Array 230A' containing 15 923 probe sets was used while it was then shifted to the 'Rat Genome 230 2.0 Array' containing 31 099 probe sets after the version upgraded.

The TGP started with employing a new normalization method, 'percellome', using an externally added spike RNA (Kanno *et al.*, 2006). This method makes it possible to express

each gene expression as copy numbers per cell (per DNA) by adding external Bacillus subtilis mRNA proportional to the DNA contents in the homogenate. In our database, raw data, per-chip normalized data as well as spike-normalized data, are usable.

Per-chip normalization (global normalization) is fundamentally based on the assumption that the total amount of mRNA is constant, and thus the change of each mRNA cannot be precisely estimated when the total transcription is drastically changed (Kanno *et al.*, 2006). In the TGP, the version of 'GeneChip' was changed and this made it impossible to make a comparison between different chips based on values with global normalization. Moreover, in the *in vitro* experiments where extremely high concentrations of the chemicals were applied, drastic changes of total mRNA were often observed. Especially, when a chemical having direct cytotoxicity to hepatocyte was applied, total mRNA was obviously decreased, and many genes, which were actually down-regulated, were estimated as apparently up-regulated because of normalization by the reduced value. In analysis of the kidney, the cell composition varies with the particular part of the organ. It was found that not only the members of expressed genes but also the total mRNA amounts differed among cortex, medulla and papilla. When region-specific genes were extracted, it was revealed that global normalization led to a biased conclusion because of the large difference in the total mRNA (Tamura *et al.*, 2006a).

Contrary to the above facts, we concluded that there is no problem in the usual analysis based on global normalization, at least *in vivo* liver data, since total RNA contents were almost unaffected, even at the toxic dose. When the ratio to control value is employed, global normalization is rather superior to spike normalization because the extra procedure introducing an additional error is avoided. In this present paper, analyses are based on global normalization.

20.2.2 Selection of Test Compounds

The TGP started with five representative hepatotoxicants, i.e. acetaminophen, isoniazid, carbon tetrachloride, phenobarbital and valproic acid. The 150 chemicals were selected and their exposure to rats have been completed – these are categorized in Figure 20.1. Although they are somewhat biased, they cover most of the therapeutic categories. At present, the drugs in the Japanese market number about 1500, including those not suitable for transcriptome analysis of the liver or kidney, e.g. drugs for dermatology or ophthalmology, drugs with almost identical structures and anti-cancer drugs whose toxicity to these organs is not a primary problem. Our aim is to create a database of representative medical compounds. In this sense, the number of 150, 10 % of total drugs, is considered to be enough for a 'textbook of toxicology'. It would be interesting to extend the project toward a structure–toxicity relationship based on the present database.

One unique feature of our project is the supply of chemicals from the member companies. These are drug candidates that were withdrawn from the various stages of the development process because of the emergence of toxicity in the liver or kidney. These are usually impossible to obtain, and are quite valuable and interesting samples. The candidates that were withdrawn mean that they were once considered to be hopeful candidates; in other words, their potential toxicity was underestimated in the early stage. Thus, they could be useful for the evaluation of our database as good model cases after our system is established.

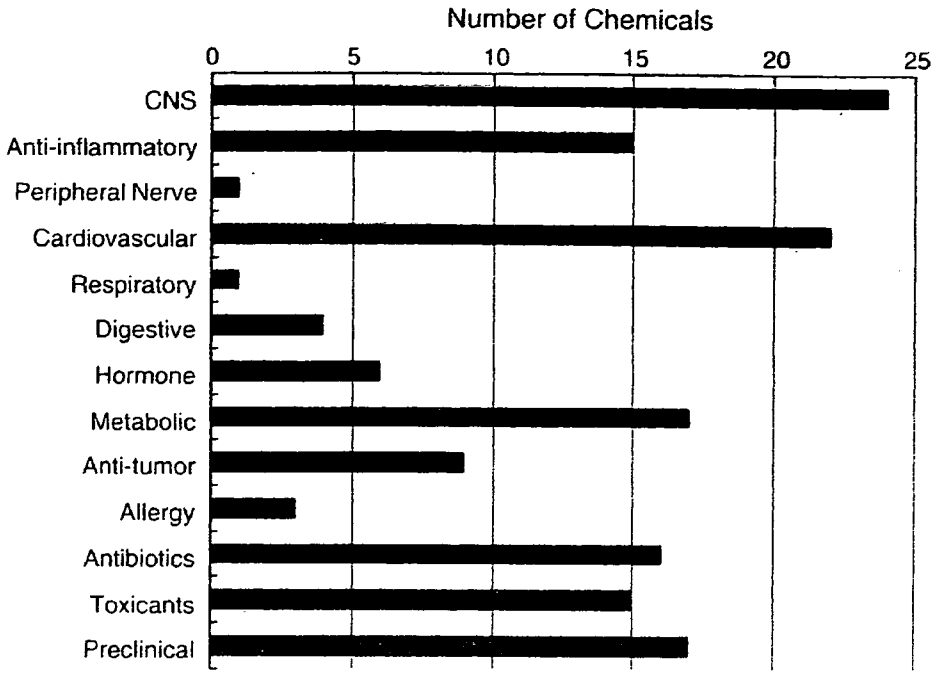


Figure 20.1 *Chemicals selected (150 total). The chemicals selected for the Toxicogenomics Project are classified by their category: CNS, drugs for central nervous system; toxicants, not medical drugs but representative chemicals with hepato- or nephrotoxicity, such as carbon tetrachloride, hexachlorobenzene, aryl alcohol, etc; preclinical, drug candidates supplied from the member companies, which were withdrawn in various stages of drug development because of hepato- or nephrotoxicity*

20.2.3 Bridging between Species

Even if the perfect prediction of hepatotoxicity is attained in the rat, this does not necessarily mean the improvement of prediction of toxicity in humans. The final goal should be the prediction in humans for drug development. There are too many problems in extrapolation of phenotype from rodent to human. However, if general toxic mechanisms or toxicological pathways are conserved over species, they would obviously be useful for bridging between animal models and clinical events. In the preliminary studies performed before the project, it was found that the gap between hepatocyte cell lines and normal liver were too wide to bridge even if they were both human, and thus normal human hepatocyte culture was employed as the only choice. At present, data collection has been completed for rat primary hepatocytes, whereas that of human hepatocytes has not been done yet, and their comparative analysis is a future subject.

In the system of analysis for the TGP database described later, the object is to overview the responsiveness of the test compounds to the biomarker gene lists and make it easy to compare them between rat and human by an automatic ortholog conversion of the marker genes. However, there have been many difficult problems pointed out in the bridging between *in vivo* and *in vitro* experiments before solving the species difference (Boess *et al.*, 2003). We also recognized this problem as the *in vitro* data accumulated. Although our final goal should be the prediction of clinical toxicity, we are now focusing on the increment of the efficacy of preclinical study as a more realistic goal. Namely, the increment of

predictivity of hepatotoxicity in rats is the first important step and thereafter the bridging between species difference is envisioned as the next step. Therefore, this topic is omitted from this present paper.

20.2.4 Enrichment of the Protocol

This point is the largest merit of the TGP and is thus discussed in detail.

20.2.4.1 Standard Protocol and Problems in Statistical Analysis of Gene Expression Data

The protocol employed in the TGP is summarized in Table 20.1. *In vivo* experiments consist of single and repeated oral administrations to rats ($N = 5$ for each group; 3 doses + vehicle control) and autopsy is done 3, 6, 9 and 24 h after the single dose and 24 h after repeated dose for 3, 7, 14 and 28 days. Data of blood biochemistry, hematology and histopathology (both liver and kidney) are obtained from all animals and the gene expression analysis in liver (also kidney in some cases) is performed in 3 out of 5 rats. *In vitro* experiments consist of rat and human hepatocytes ($N = 2$ for each group; 3 concentrations + vehicle control) and cell harvest is done at 2, 8 and 24 h after exposure.

Table 20.1 *Standard protocols employed in the Toxicogenomics Project*

In vivo	
Animal	Sprague–Dawley rat 6 week old $N = 5$ for each group
Vehicle	0.5 % methylcellulose or corn oil
Dose	Low, middle, high (1:3:10)
Route	Oral (intravenous in a few cases)
Sacrifice	3, 6, 9 and 24 h after a single administration 24 h after the last dose of repeated administration for 3, 7, 14 and 28 days
Sampling	Liver, kidney, plasma
GeneChip analysis	$N = 3$
Items examined	Histopathology: liver and kidney Body weight, organ weight (liver and kidney), food consumption Biochemistry, hematology: 37 standard items
In vitro: rat	
Animal	Sprague–Dawley rat 6 week old
Cell	Hepatocyte isolated by collagenase digestion
Vehicle	Culture medium or DMSO
Concentration	Low, middle, high (1:5:25)
Treatment	2, 8 and 24 h
GeneChip analysis	Duplicate
Items examined	Cell biability (LDH release and DNA contents)
In vitro: human	
Cell	Human frozen hepatocytes
Vehicle	Culture medium or DMSO
Concentration	Low, middle, high (1:5:25, low is omitted in some cases)
Treatment	2, 8 and 24 h
GeneChip analysis	Duplicate
Items examined	Cell biability (LDH release and DNA contents)

We believe that no other database contains data with such an enriched protocol in the world. Especially the fact that dose-dependency with $N = 3$ for each time point can be estimated is very powerful in erasing 'noises' inevitably associated with the statistical analysis of vast numbers of measurements. Although the price of 'GeneChip' is decreasing as it becomes more popular, the cost of the experiments using this device is still so high that either the sample numbers in a group, time points, or dose levels need to be reduced. However, the reduction of the dose levels ruins the analysis. The analysis of microarray data is usually performed by multivariate statistical techniques such as hierarchical clustering, *k*-means clustering, self-organizing map or principal component analysis (Kaminski and Friedman, 2002; Draghici, 2003). When the data sets are divided into 'positive' and 'negative' by any definition, discriminant analysis such as PAM (Tibshirani *et al.*, 2002) or SVM (Brown *et al.*, 2000) can be used. However, in any case, no confident results are obtained unless the size of the gene list is reduced to a reasonable level.

Suppose only one 'toxic' dose was tested and compared with its control. When using the A chip, 15 923 pairs should be made and 31 099 pairs for the v.2.0 chip. Even when the *p* value is set to a very low value, too many 'significant' differences without biological significance are obtained. On the other hand, when one does not want to miss a certain gene with high significance in terms of biology or toxicology, the *p* value cannot be lowered in the experiments with such small *N* values. It has been pointed out that application of the usual biostatistics is difficult when the evaluation is done with very small *N* values against vast numbers of measurements such as whole genes. Especially when the data contain biologically based variations, no improvement is expected by the use of any sophisticated statistical technique. When data with multiple doses or time point are available, however, the extraction of biologically meaningful changes is easier.

20.2.4.2 Example 1: The Case of Omeprazole

Let's refer to the actual data. The first one is omeprazole, which was administered to rat at 100, 300, and 1000mg/kg and analyzed with the A chip (15 923 probe sets). In the TGP database, there are three dose levels with vehicle control for each of four time points in single and repeated doses, respectively. When comparison between control and treated rats is made in each point, it will be 380 740 pairs. When an uncorrected Student's *t*-test is applied, 36 883 pairs of 'significant at $p < 0.05$' are obtained. No one would like to perform a precise analysis of these probe sets and it is practically impossible. It is common sense in statistics that this number is overestimated and some correction is needed. It is necessary to erase the accidental difference not related to the drug effect, but the usual statistics only tells you to reduce the *p* value based on some criteria. A simple reduction of the *p* value to 0.001 results in the reduction of the 'significant' pair numbers to 1639, which is still a quite large number, but the main question here is whether this has elucidated the really biological meaningful differences. Regrettably, the answer is no, in most cases.

Suppose 'genes significantly changed by 100 mg/kg omeprazole at any time point' were extracted by the *t*-test without correction. Figure 20.2(a) shows an arbitrarily selected probe set, 13 938 56 ×. At $p < 0.05$, this gene is considered to be up-regulated at 3 h whereas it is down-regulated at 9 h after administration. However, when all of the data are reviewed (Figure 20.2(b)), it is quite easy to conclude that no biologically meaningful change is caused by the drug, since there is no dose- or time-dependency. This type of extraction only creates a 'heap' of 'junky' genes.

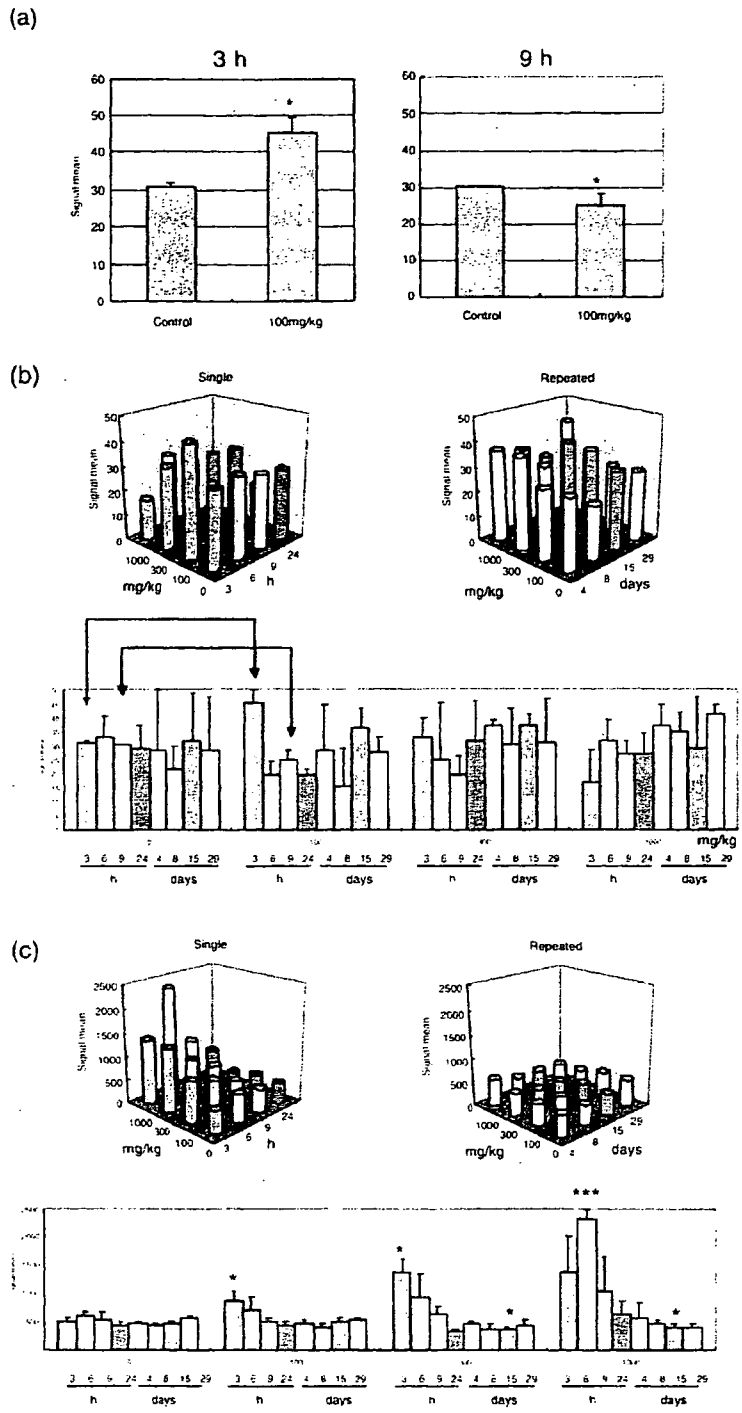


Figure 20.2 *Effects of omeprazole on gene expression. (a) Effects of a single oral dose of omeprazole (100 mg/kg) on the expression of an arbitrarily selected probe set, 1 393 856× at 3 and 9 h after dosing. (b) Effects of single and repeated oral doses of omeprazole (100, 300, 1000 mg/kg) on the expression of 1 393 856 ×. The upper panels show the 3-dimensional graphs of single (left) and repeated (right) dosing, while the lower panel shows the 2-dimensional graph with error bars. The arrows indicate where 'statistical significance' was noted in Figure 20.2(a). (c) Effects of single and repeated oral dose of omeprazole (100, 300, 1000 mg/kg) on the expression of heme oxygenase-1 (1370080_at). The upper panels show the 3-dimensional graphs of single (left) and repeated (right) dosing. The lower panel shows the 2-dimensional graph with error bars. Expression data were normalized by mean value and multiplied by 500, and expressed as the mean of N = 3 with SD where indicated. The asterisks indicate 'statistical significance' by the uncorrected Student's t-test at* $*p < 0.05$ *and* $***p < 0.001$

Next, attention is paid to heme oxygenase-1, which is known to play important roles in cellular stress response (Figure 20.2(c)). These data obviously indicate that omeprazole potently and dose-dependently induces this gene at 3 h after dosing. Here, if gene extraction is performed by using only the highest dose (1000 mg/kg), only the change at 6 h should be noted. An even worse point is that one would conclude that this important gene is not responding to 1000 mg/kg omeprazole if extraction is performed without using the data of 6 h. Looking at Figure 20.2(c), most researchers would agree that omeprazole induces the expression of heme oxygenase-1. However, it is highly possible that omeprazole is regarded as a negative inducer of this gene by experiments with reduced data points and restricted statistics.

20.2.4.3 Example 2: Age-Related Difference in Toxicity

Before the start of the TGP, there was an argument regarding the age of rats. One asserted that 6-week old rats should be used because this age is recommended in the standard toxicity tests, whereas the other claimed that confident data with small variations would be only obtained from mature animals not younger than 10 weeks. Before reaching a conclusion to use 6-week old rats, a study to compare 6- and 12-week rats was performed regarding the sensitivity of hepatotoxicity.

Acetaminophen, isoniazid and carbon tetrachloride were selected as representative hepatotoxicants and the sensitivity to those were compared between 6- and 12-week old rats by a single administration protocol. Although the latter two showed no age-related difference, acetaminophen was found to be more toxic in 12-week old rats than in 6-week old rats. The causal factors were suggested to be the higher expression of CYP3A13 that produces an active metabolite and/or the lower expression of a subtype of glutathione transferase (Morishita *et al.*, 2006). The most interesting findings in this study were the following points. In order to compare gene expression changes at 24 h after dosing where pathological changes emerged, differentially expressed genes between 6- and 12-week old rats were extracted by statistics. In the usual course of investigation, one would try to attribute the difference in toxicity to the expression level of these genes by using correlation or discriminant analysis. However, a precise review of these genes revealed other features. Among the stress-responsive genes, many genes showed age-related difference not in the extent but in the time course.

Figure 20.3 shows again a representative stress responsive gene, heme oxygenase-1. In 6-week old rats, the peak of induction by acetaminophen appeared at 9 h or earlier and the expression returned to basal level at 24 h, whereas the peak was later than 9 h in 12-week old rats. If the observation were done at 24 h only, the expression of heme oxygenase-1 would have been judged as 'yes' in 12-week old rats and 'no' in 6-week old rats by an 'all-or-none' manner. However, the actual response was 'yes' in both cases and the difference was present in the response time. Other information from Figure 20.3 is that a threshold dose of acetaminophen exists in the induction of at least this gene. This is reasonable, considering the widely accepted toxic mechanism of this drug, i.e. the hepatocyte damage does not occur when glutathione is not depleted and the detoxification system for active metabolite is active (James *et al.*, 2003). In the present case, the induction of this gene would be undetectable unless the test with the highest dose was performed.

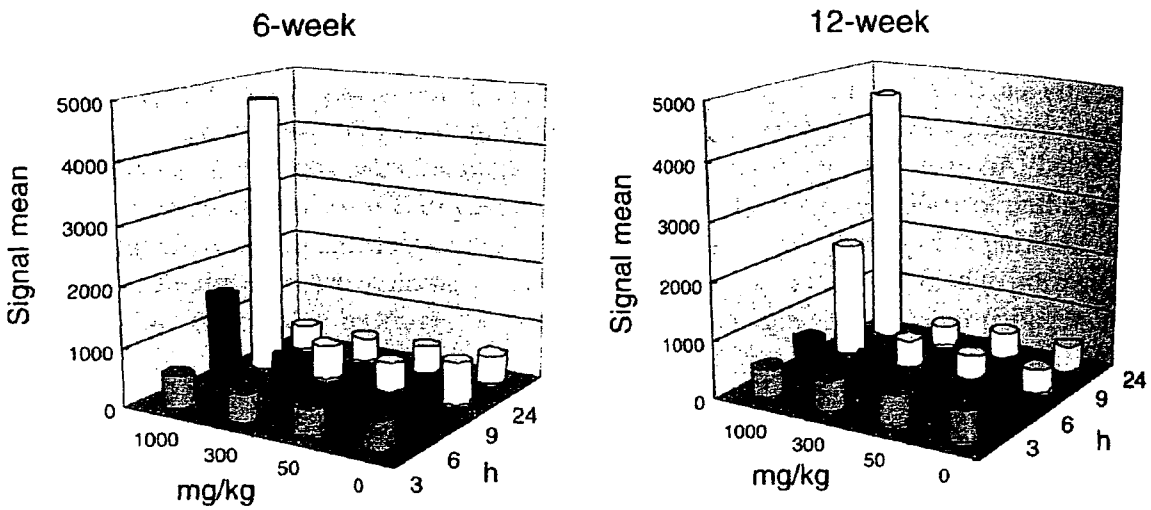


Figure 20.3 *Effects of a single oral dose of acetaminophen (50, 300, 1000 mg/kg) on the expression of heme oxygenase-1 (1370080_at). The left and right panels show the data of 6-week old and 12-week old rats, respectively. Expression data were normalized by mean value and multiplied by 500, and expressed as the mean of N = 3*

20.2.4.4 Advantage of Multi-Time, Multi-Dose Protocol

One would expect that the sensitivity of transcriptome analysis should be always higher than any traditional toxicological technique. Who wants to do additional, quite expensive and tedious tests when obvious toxicological phenotypes are already obtained? However, the previous example tells us that this hope is too optimistic, especially when a toxicological threshold may exist in a certain step of a sequence of gene expression, as in phenotypes. There would be two alternative answers. One is that prediction of toxicity by toxicogenomics technology is only possible in a chronological manner where the administration of toxic doses is essential. In this case, the prediction would be 'Keeping this dosage will cause that phenotype, etc'. The other is that the toxicity of overdosage with apparent threshold can be indirectly predicted even at the low dose by gene expression changes somewhere in the toxicological pathway if threshold does not exist in that step. This point will be discussed later.

Figure 20.4 depicts a schematic expression of time- and dose-dependencies of gene expression changes. Needless to say, the conclusions drawn from using multiple components with different time- and dose-dependencies must differ among the cases, where a single observation each is made at time point A or B, or at dose level X or Y. In the case of the usual biology test focusing on a particular target, it would be possible to set an appropriate time and dose in a preliminary study. However, toxicogenomics is designed to observe any gene expression changes reflecting 'any toxic phenomena in the future' one chip at a time, and so it is practically impossible. In order to 'mine' the data with biological significance (this does not always coincide with statistical significance), data with multiple time and dose are considered to be essential. Any sophisticated statistical procedure cannot create anything from no data.

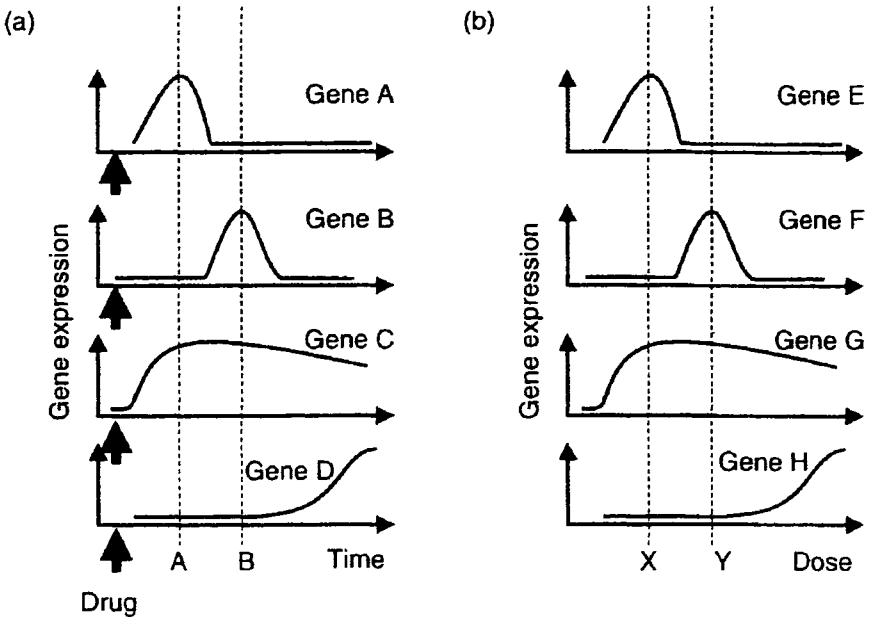


Figure 20.4 *Schematic expressions of (a) time- and (b) dose-dependencies of gene expression changes. (a) Some of the genes might be up-regulated by the administration of a drug (indicated by arrows) either transiently at an early time point (A) or at a later time point (B), or continuously with early (C) or later onsets (D). If observations are made either at the time point A or B, the results would be quite different, i.e. up-regulation is observed in A and C at time point A, whereas up-regulation occurs in B and C at point B and nothing is detectable in D. (b) Similar to the above, some of the genes might be up-regulated by the administration of a drug with either bell-shaped (E, F) or sigmoidal (G, H) dose-dependency. If observations are made at the concentrations of either X or Y, the results would be quite different, i.e. up-regulation in E and G at concentration X, whereas up-regulation in F and G at Y with nothing detectable in H*

Even with such enriched data stored in our database, the procedure as to how to efficiently withdraw significant genes is immature. A statistical test covering dose-dependency, Williams test, for example, is available, but it does not always work efficiently. In general, three for each group is too small for any statistical analysis. However, the cost effectiveness is still questionable when the number is increased to, say, five. In order to make the statistical analysis applied to over 10 000 measures meaningful, N should be increased to a similar order, which is impractical in the biological data. In our experience, too strict statistics should not be applied for extracting significantly mobilized genes, in order not to overlook important genes with biological variance. We use various properly applied approaches case-by-case, e.g. ANOVA with a relatively large p, followed by proper filtering (elimination of the genes with significance in low dose only, or whose expression were inversely correlated with dose, etc.) or selection without statistics, based on the value of ratio to control. In any case, all of the data of extracted genes down to a reasonable number are stored in the database and so at anytime can refer their dose- or time-dependencies.

20.2.4.5 New Knowledge from Accumulated Data

When vast amounts of data are accumulated in the database, an interesting thing emerges by simple alignment of the data. Figure 20. 5(a) presents a summary of the vehicle control

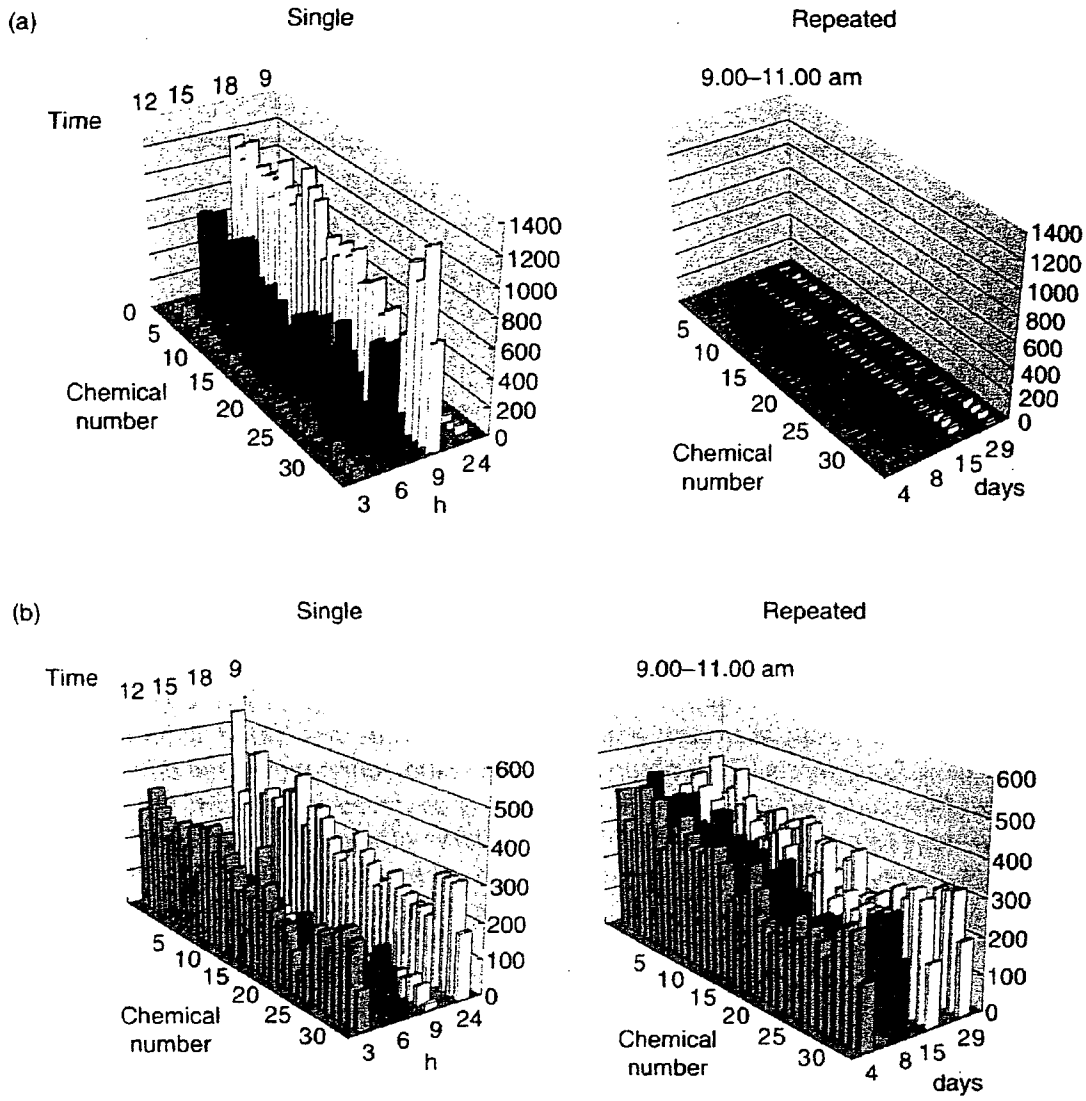


Figure 20.5 *Circadian changes in the expression of (a) D the site albumin promoter binding protein, 1387874 and (b) the aryl hydrocarbon receptor nuclear translocater-like, 1370510 in vehicle controls for the early 35 chemicals in the database tested by using the RAE 230A chip. Left panels – circadian changes of the expression within a day. Sacrifice of the animals in single-dose experiments is done 3, 6, 9 and 24 h after oral administrations in the morning. They thus correspond to about 12.00 midday, 15.00 pm, 18.00 pm and 9.00 am the next day, respectively. Right panels – in repeated administration, sacrifice is done in the morning at around 9:00 am to 11:00 am, and thus the expression level is constant to the value in the morning through all of the chemicals and days. It is also obvious that expression does not change with age*

group of single- and repeated-dose experiments. The left panel shows the expression level of a representative circadian gene, D site albumin promoter binding protein, for 3, 6, 9 and 24 h data of the vehicle controls from the first 35 chemicals tested with an old A chip. As the administration is done in the morning, this pattern shows the reproducible increase of expression in the afternoon. The right panel shows its expression level in the repeated dosing for 3, 7, 14 and 28 days. As the autopsy was done 24 h after the last dose, i.e. in the morning, the expression level appears uniformly low. On the other hand, there are genes

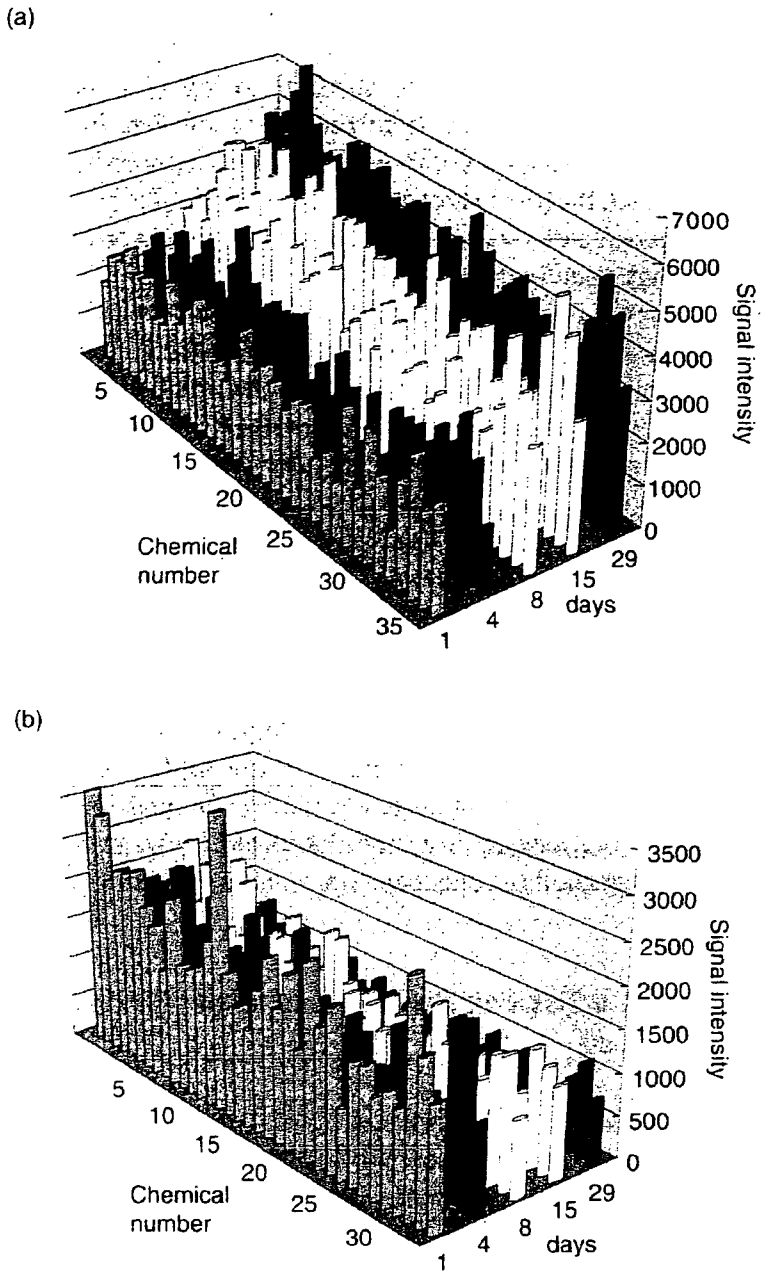


Figure 20.6 *Expression levels of (a) steroid delta-isomerase 3 and (b) the hemoglobin beta chain complex in vehicle controls for the early 35 chemicals in the database tested by using the RAE 230A chip*

with reversed circadian rhythm such as ARNT-like (Figure 20.5(b)), i.e. high expression level at noon and it goes down towards the night. The field of circadian gene expression has recently attracted attention (Ueda *et al.*, 2004). Using our TGP database, many genes with circadian rhythm, as well as potential drugs that affect the rhythm, can be easily extracted.

Reviewing the vehicle control data of repeated dosing experiments, we noticed a group of genes whose expression level changed with age. Figures 20.6(a) and 20.6(b) show the expression levels of steroid delta-isomerase 3 and hemoglobin beta chain complex, respectively. The former increased, while the latter decreased with age, possibly reflecting

a physiological increase in steroid hormone synthesis and a decrease in extramedullary hematopoiesis with age.

In TGP, either 0.5 % methylcellulose or corn oil works as vehicle, according to the solubility of the test drug. As a lot of data of both vehicles were being accumulated, the extraction of vehicle effects became possible. As we have already published (Takashima *et al.*, 2006), it was revealed that the extent and/or time course of expression of genes related to lipid metabolism was affected by corn oil, since rats received orally a high caloric intake in the morning when they usually do not eat.

These genes with reproducible expression patterns are now utilized for quality control of each expression data in the TGP. This could be applied for evaluation of data obtained in a different platform in the future. There are also many probe sets with quite poor reproducibility or large variance. If the cause of their variation is attributed to certain factors, like animal treatment or laboratory circumstances, such genes are in turn useful tools to evaluate inter-laboratory differences. It might be convenient to make a list of 'useless' genes that are excluded at the beginning in order to facilitate the analysis. However, care should be taken to exclude any genes that are absent in samples in the database even after a vast amount of data is accumulated, at least insofar as 'toxicity' is concerned. It is always possible that expression of such a gene is uniquely induced by a new drug.

Figure 20.7 shows the overview of the expression of the TNFRSF16 associated protein 1 in the first 35 chemicals analyzed with the A chip. Only thioacetamide and methapyrilene induced its expression after repeated administrations. It would be interesting to investigate the toxicological mechanism of these non-genotoxic carcinogens making this observation as a clue. The point here is that the simple accumulation of the vast amount of data is scientifically valuable and it is important to collect any observable changes as precisely as possible.

20.3 Construction of a Toxicity Prediction System Based on the TGP Database

The TGP system consists of basically three parts, i.e. the database itself that stores gene expression data with related pathology (scoring and the photo of HE staining), hematology, blood biochemistry and chemical information, the data analysis system that consists of the tools for up- and download of data, clustering, discriminant analysis, principal component analysis, gene- or compound list manager, etc. and the prediction system that is used when the expression data are uploaded.

In the summer of 2006, the final form of these systems has become operational. In the TGP database, *in vivo* data of 24 000 rats, expression data of about 24 000 'GeneChips' corresponding to ca. 700 000 000 probe sets, 2 880 000 measured test items, the data of 48 000 pathology specimens and various related information and reports are to be stored in their final form. In order to pick up useful data for toxicologically meaningful analysis from such a large scale of data, an efficient, toxicologist-friendly system is essential.

20.3.1 Analysis System

Let's see the previous omeprazole case again. The first one is pathology. In the single dose experiment, the significant and dose-dependent change was periportal eosinophility

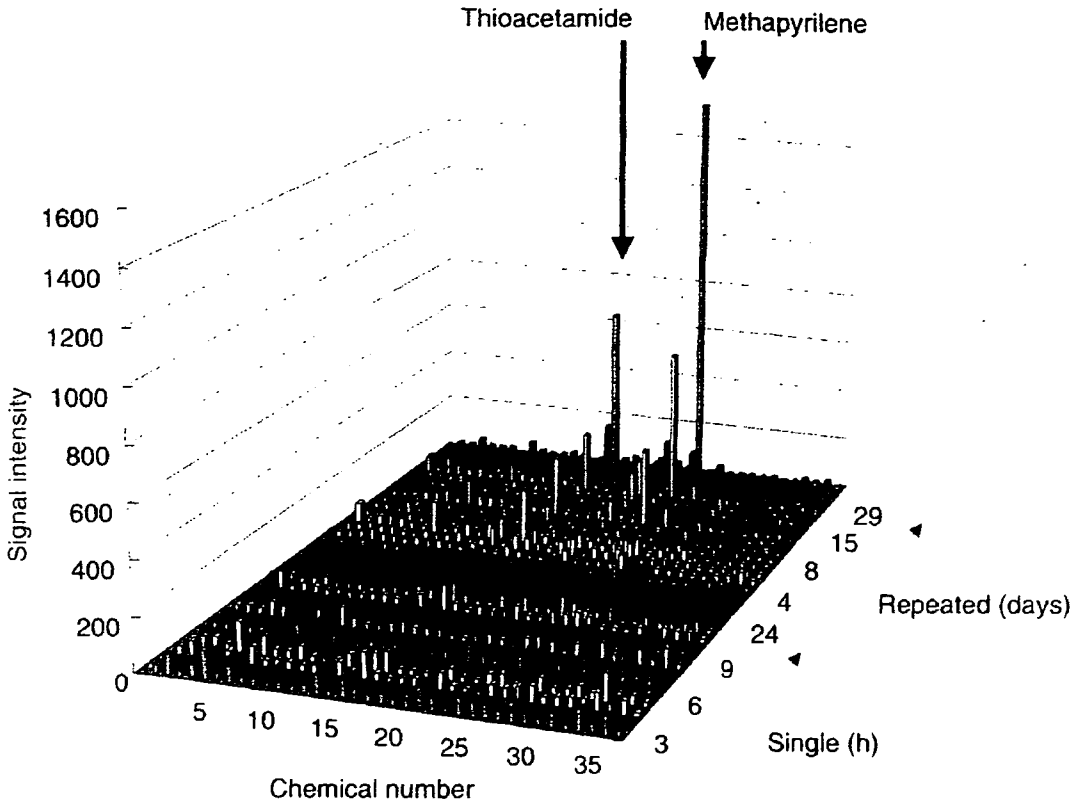


Figure 20.7 *Expression level of the TNFRSF16 associated protein 1 in the first 35 chemicals analyzed by using the RAE 230A chip. This gene was specifically up-regulated by repeated administration of thioacetamide (indicated by the arrow around the chemical number 17) and methapyrilene (indicated by the arrow around the chemical number 25). Expression data were normalized by mean value and multiplied by 500, and expressed as the mean of N = 3. On the time axis, sets of the data of control, low, middle and high doses are aligned in the order of 3, 6, 9 and 24 h after single doses, and 3, 7, 14 and 28 days of repeated doses*

observed at 6 and 9 h. In the repeated ones, centrilobular hypertrophy was evident with a peak at 2 weeks. Simultaneously measured test items were 43 in total, and 22 items showed some changes. As it is quite tedious to check each individually, the extent of the change to control value in each measure was converted into a semi-quantitative heat-map and depicted as Figure 20.8(a). It can be seen that omeprazole did not induce serious toxic changes in a single dose, whereas obvious hepatic hypertrophy, anemia and some mobilization in plasma lipid emerged by repeated administration.

The next one is the gene expression change in the liver. For each gene (probe sets), a graph of 4×4 matrix with $N = 3$ for each lattice can be drawn for single and repeated experiments. As it is again quite difficult to interpret the results consisting of more than 15 000 probe sets, checking one by one, a similar heat-map was depicted by a semi-quantitative conversion of the dose–response at each time point (Figure 20.8(b)). This is for the example of glutathione reductase. It is obvious that this gene was dose- and time-dependently up-regulated toward 24 h in the single dose, whereas the extent kept decreasing as the administration continued. Since glutathione reductase is known to involve oxidative stress responses, other genes

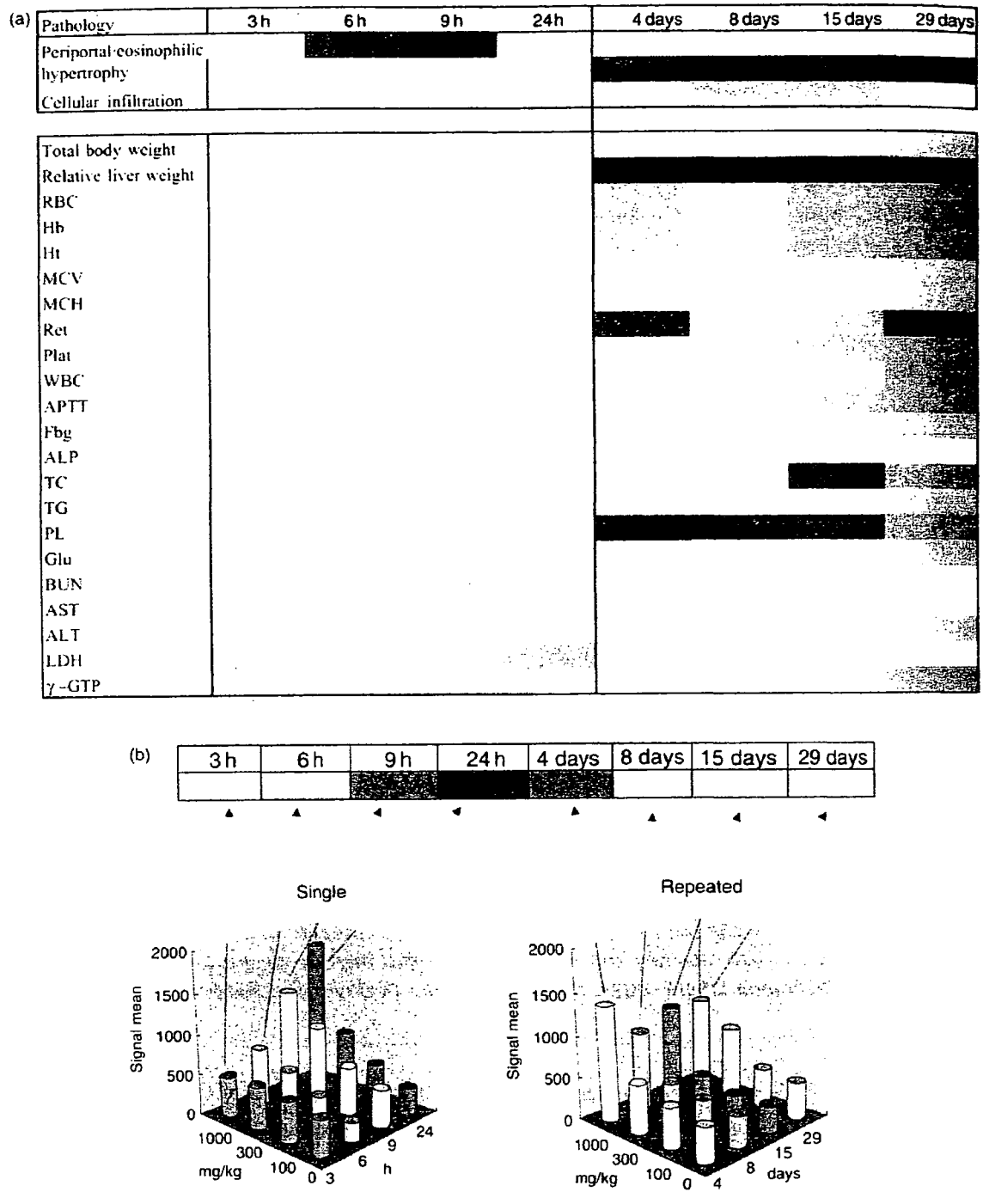
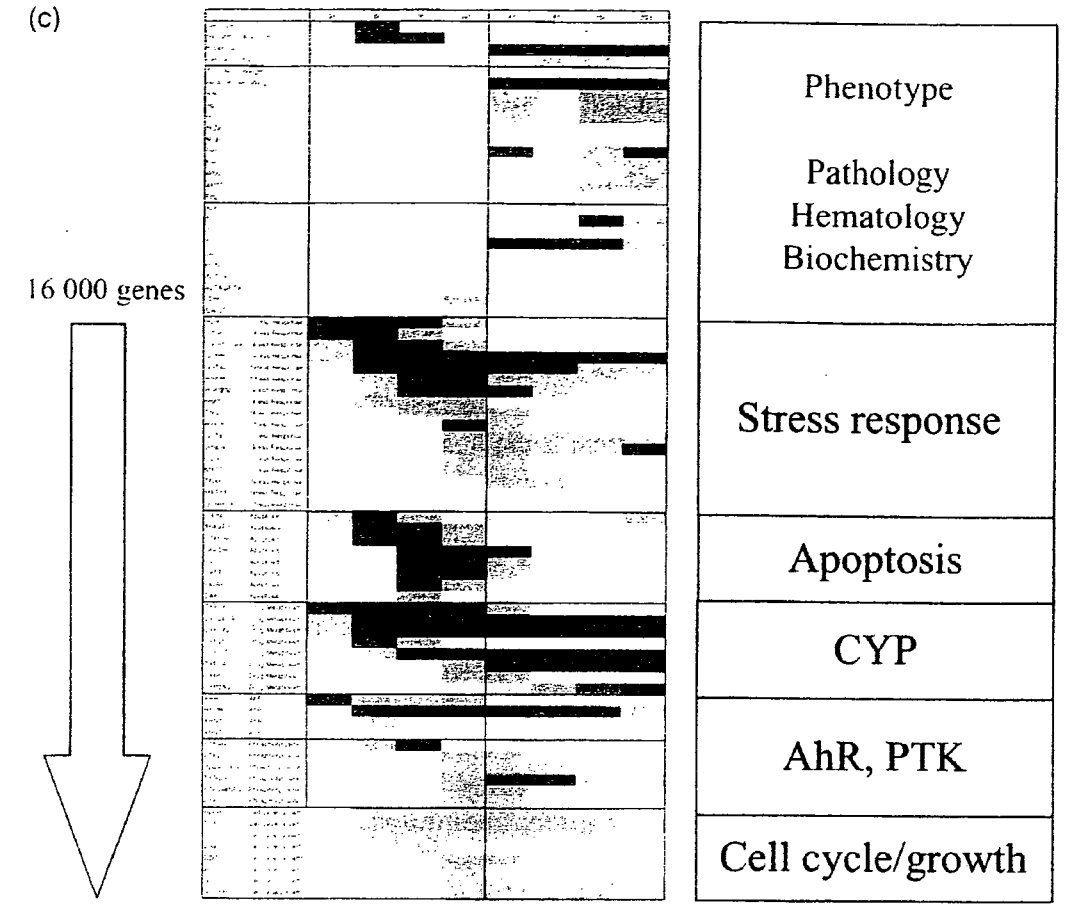


Figure 20.8 *Schematic representation of an image of data stored in the TGP database. (a) Heat-map of phenotype induced by omeprazole administered to rats as an example. The upper panel shows the pathological findings while the lower panel shows the changes in organ/body weight, plasma biochemistry and hematology. The actual data table is much larger than this, but the items without significant change are omitted and the data are converted to a semi-quantitative heat-map. (b) Heat-map conversion of the expression changes of glutathione reductase, 1369061_at. This way makes it convenient to overview time- and dose-dependent changes at a glance. (c) Heat-map expression of the gene expression changes together with the phenotype induced by omeprazole. Genes were categorized by their function and aligned by the order of the time when the first change appeared. This panel continues far down below. (d) The alignment of the panels of the chemicals prepared by the way described above. Practically, it continues in both directions of horizontal and vertical*



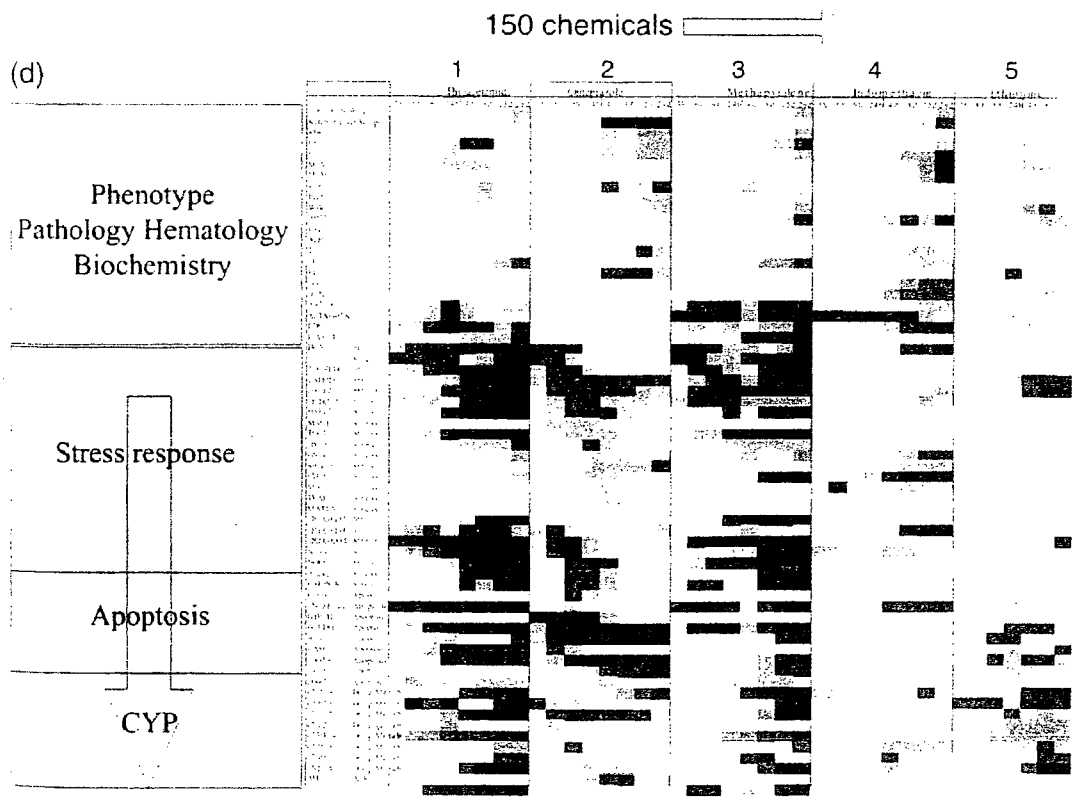


Figure 20.8 (Continued)

belonging to this category were examined and aligned in the order of their responding time to overview the sequence of chronological expression of the functional genes. Continuing this step for other categories, it produces a long heat-map as shown in Figure 20.8(c), which actually continues far down below. It might be possible to hypothesize a cascade of gene expression related to the effects of omeprazole on the liver, but the whole body of data is too large to reach an acceptable conclusion without the aid of computerized pathway tools.

Moreover, 150 in such a heat-map are lining horizontally (Figure 20.8(d)). The main problem here is how to extract the genes whose expression changes are correlated with a certain phenotype or potential toxicity by using the bioinformatics technique.

In the case where a target or phenotype is clear, such as PPAR agonists, the extraction of significant genes is relatively easy (Tamura *et al.*, 2006b). However, it is quite difficult to extract genes related to 'toxicity' with unknown mechanisms. For example, suppose one would try to extract genes related to above 'hypertrophy' caused by omeprazole. It would be unsuccessful if one extracts the commonly changed genes among the compounds that caused hypertrophy. This is not a matter of statistical technique but of pathology. First, it is quite rare that a certain pathological phenotype is attributed to a single toxicological mechanism. Secondly, the quantitativity of pathology scoring is poor. Thirdly, when the severity of a phenotype increases or decreases, its grading score does not always increase or decrease, but the name of the diagnosis changes. One cannot avoid this problem by including the diagnosis expressing the upper or lower grade, since the name does not necessarily express the severity of the phenotype alone. 'Phenotype anchoring' is a challenging issue in the toxicogenomics field (Moggs *et al.*, 2004) and it is also under investigation in our TGP. It is also expected that progress in 'toxicogenomics-oriented histopathology' should be an important field in toxicology.

We have therefore started with anchoring to clarify indicators in parallel. The TGP database contains representative drugs and hepatotoxicants, but it does not necessarily mean that their toxicological mechanisms are representative with respect to molecular biology. We then decided to perform additional single-dose experiments using compounds that possess clear molecular targets in order to investigate the relationship between molecular toxicity and gene expression changes. They included protein synthesis inhibitors, compounds related to infection/inflammation (lipopolysaccharide, TNFα, galactosamine), ER-stressor, and cytoskeleton disruptors. The results of phorone, a glutathione depletor, are presented.

As described above, hepatotoxicity of acetaminophen is due to excess production of active metabolite over the detoxification capacity of intracellular glutathione (James *et al.*, 2003). Therefore, any drugs that have a potential to deplete hepatocyte glutathione risk causing acetaminophen-type hepatotoxicity with overdosage. In a previous report, a list of marker genes for glutathione depletion was extracted using BSO, a glutathione biosynthesis inhibitor (Kiyosawa *et al.*, 2004). Phorone was considered to be superior to BSO as a model system, since the mechanism of glutathione depletion is similar to that of acetaminophen-type hepatotoxicants, i.e. it covalently binds to glutathione and is excreted from the cell.

Phorone, at 40, 120 or 400 mg/kg was administered according to the same protocol as the regular single-dose experiments and the measurement of glutathione contents was performed in addition to the regular tests. Phorone caused a marked but transient depletion of glutathione with maximal depletion occurring at 3 h; then it recovered and showed

an increase of glutathione at 24 h as a rebound. A significant increase in plasma AST was observed at 400 mg/kg, indicating hepatotoxicity. Expression data at this point were excluded from the work-up in order not to extract the expression changes secondary to hepatic injury. In the next step, genes whose expression were inversely correlated with hepatic glutathione contents for each rat were statistically extracted and filtered to get 130 probe sets. Principal component analysis of the chemicals stored in the database using these sets revealed that chemicals with a risk of glutathione depletion, such as bromobenzene and coumarin, in addition to acetamineophen, were clearly separated from other chemicals or controls toward the direction of PC1, suggesting that the list was a useful as 'marker gene list for risk assessment of glutathione depletion' (Kiyosawa *et al.*, in press).

Our present strategy is to prepare biomarker gene lists that are related to certain toxicological phenotypes, pathways, or any biologically meaningful factors, as many as possible using various procedures.

20.3.2 Toxicity Prediction System Based on Biomarker Gene Lists

In the general terminology, 'biomarker' is designated as one or a few biological measures quantitatively reflecting a certain biological change. However, this does not fit for the 'biomarker gene' in toxicogenomics. As discussed above, toxicology based on the transcriptome has various problems, such as poor statistics because of small *N*s compared with genes, the requirement to make the beta-error small, toxicity with uncertain time- or dose-dependency, etc. Although the quantitativity of microarrays has greatly improved in recent years, its quantitativity and reproducibility cannot be superior to the enzymes in serum or the metabolites in urine when a labile object, mRNA, is measured. It is thus dangerous to make a decision based on one or few marker genes, and it would be effective to use relatively large numbers of genes as a whole in order to make the assessment robust. A comprehensible example is shown in Figures 20.2(c) and 20.7(c). If heme oxygenase-1 alone is employed as a 'stress marker', the potential to overlook omeprazole must be quite high, but it should go down if a 'stress-responsive gene list' is employed. In this connection, the TGP is now trying to substantialize such gene lists.

When an assessment or prediction of toxicity is made by a list of multiple measures, it becomes necessary to summarize or quantify these measurements. Ideally, the quantification process should be optimized for each marker gene list. However this is practically difficult and thus a uniform system has to be created. In the TGP, a new scoring system was developed in one trial (Kiyosawa *et al.*, 2006). The score is calculated based on the ratio to control value (log 2) for each gene in the marker list and expressed as a heat-map. This scoring system has made it easy to overview the assessments of a target compound against many marker lists, or to overview the assessments of many compounds against a particular marker list. However, there are some problems in this system, i.e. the score is biased when the list contains a gene whose expression change is extremely large (e.g. CYP1A1) and changes are canceled when up- and down-regulated genes co-exist in the list. Therefore, another scoring system, e.g. effect size (the absolute value of the difference between means divided by the standard deviation) is also available in the TGP system.

Principal component analysis is a quite convenient tool to make a qualitative classification of compounds against a list of genes. As a prediction system, however, some quantitative

data would be favorable for the final output. Thus in our system, the following functions are added, i.e. when the user specifies a principal component with high contribution, the compounds are sorted by the value and the genes with large Eigenvector value are easily obtained. This gives an idea where the relative position of the test drug locates among the ones in the database, and suggests a candidate gene list for further investigation.

When a phenotype that can be judged as positive/negative is available, discriminant analysis is known to be powerful (Porter *et al.*, 2003), and prediction analysis of microarray (PAM) (Tibshirani *et al.*, 2002) has been firstly employed in the TGP. By a semi-automatic system of training and validation, the efficiency improves for the creation of discriminators. As above, the system exhibiting the prediction by PAM as quantitative scores (to show the relative position of a test drug among chemicals in the database) is under consideration. The TGP also includes the support vector machine (SVM) (Brown *et al.*, 2000) in the system.

Although the present system has not come to completion, the following picture of TGP use has emerged (Figure 20.9).

You have a candidate drug, X, which was administered to rats and 'GeneChip' data of the liver were obtained 24 h after dosing. The data are up-loaded to the TGP system and the marker viewer is activated to overview all the biomarker gene lists stored in the database. Alarms are noted for several markers (Figure 20.9(a)). Among them, one marker, M1, is selected, as this is highly related to the toxicological phenotype F1, if repeatedly administered. When PCA is performed using these marker genes, X is clearly separated in the direction of PC1 (Figure 20.9(b)). As the contribution of PC1 is found to be high, compounds are sorted by PC1 and their order is Y (high dose) > X > Z (high dose) > Y (middle dose) > ...., etc. From the analysis, it is predicted that X causes phenotype F1 when repeatedly dosed, and it requires a higher dose than Y but lower than Z. The gene list with a high Eigenvector is harvested for further analysis of the toxicological pathway. There is also another point. It is known that X is pharmacologically similar to compounds P and Q that are stored in the database. It is also known that P and Q cause phenotype F2 and actually that a marker gene list M2, predicting F2, exists in the database, whereas no alarm was noticed in the first survey. Then all the compounds are overviewed against M2. It is obvious that not only X but also P and Q show low scores because most of the marker genes in the list show transient expression changes and thus F2 is non-predictable using M2 at 24 h (Figure 20.9(d)). Therefore, it is suggested that an additional early time point is necessary to judge whether X has a similar property to P and Q.

This is just a simulation. The point here is that the prediction system of the TGP is not a simple output of a probability like a 'weather-forecasting system', but a supply of knowledge with suggestions for further investigation. As everybody knows, there is no drug without side effects. If a prediction like 'this drug is safe' appears, this must be a lie. Toxicologists do not want such a prediction, but want the information such as, 'What kind of phenotype would appear in what dose level and what are the toxicological mechanisms involved?'.

20.4 The Image of Toxicity Testing after the TGP Database is Established

At present, we say that the usage of the TGP database/prediction system is in the following condition. In the quite early stage of drug development, the database is used to select a lead compound among candidates. As the full-scale toxicity test is quite costly, safety assessment