Development of an Approach for *Ab Initio* Estimation of Compound-Induced Liver Injury Based on Global Gene Transcriptional Profiles

Xudong Dai
xudong_dai@merck.com

Yudong D. He
yudong.he@merck.com

Hongyue Dai
Hongyue.dai@merck.com

Pek Y. Lum
pek.lum@merck.com

Christopher J. Roberts
Robchris@merck.com

Jeffrey F. Waring
Jeff.waring@abbott.com

Roger G. Ulrich
Roger.ulrich@merck.com

1 Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109 USA
2 Abbott Laboratories, Abbott Park, IL 60064 USA

Abstract

Toxicity is a major cause of failure in drug development. A toxicogenomic approach may provide a powerful tool for better assessing the potential toxicity of drug candidates. Several approaches have been reported for predicting hepatotoxicity based on reference compounds with well-studied toxicity mechanisms. We developed a new approach for assessing compound-induced liver injury without prior knowledge of a compound's mechanism of toxicity. Using samples from rodents treated with 49 known liver toxins and 10 compounds without known liver toxicity, we derived a hepatotoxicity score as a single quantitative measurement for assessing the degree of induced liver damage. Combining the sensitivity of the hepatotoxicity score and the power of a machine learning algorithm, we then built a model to predict compound-induced liver injury based on 212 expression profiles. As estimated in an independent data set of 54 expression profiles, the built model predicted compound-induced liver damage with 90.9% sensitivity and 88.4% specificity. Our findings illustrate the feasibility of *ab initio* estimation of liver toxicity based on transcriptional profiles.

Keywords: hepatotoxicity, *ab initio*, machine learning, neural network, toxicogenomics

1 Introduction

Toxicity is a major cause of failure during drug development [8, 12], particularly for compounds in the preclinical stage [12, 13]. As the main site for xenobiotic metabolism, the liver is the leading toxicological target organ and thus of paramount importance to better assess and predict the potential hepatotoxicity of drug candidates at earlier development stages. Global expression profiling based on microarray and other platforms provides a powerful tool to address this issue [14]. Microarrays are increasingly used to identify discrete gene sets associated with a specific toxic response [9, 10, 18, 20] or to survey genes affected by compounds with known mechanisms [2, 4, 10, 12]. The majority of reported approaches focus on illuminating potential mechanisms of toxicity for unknown compounds using signature gene sets identified from the expression profiles of well-studied reference compounds. This strategy depends on the accuracy of the toxicity mechanism assigned to the reference compounds. It is possible, for example, that multiple molecular mechanisms underlie the toxic action of a single compound, and that different mechanisms may dominate at different dose levels. Therefore, it would be beneficial to develop an approach for *ab initio* assessment of compound-induced liver damage from expression profiles without prior knowledge about toxicity mechanism.
Two broad categories of analytical methods have been used in the majority of reported toxicogenomic studies [19]. The first category includes unsupervised clustering approaches, which allow compounds to be clustered based on similarities across sets of genes differentially regulated by various treatments. Toxicity of compounds within the same cluster is then predicted using "guilt by association." The second category is referred to as supervised clustering or classification approaches. Different from the concept of classification in the data mining field, this classification approach in principle resemble clustering based on a set of selected genes. Specifically, a set of genes commonly regulated by a compound or a set of compounds associated with a class of well-studied toxicity (e.g., DNA damage or induction of apoptosis) is identified as the reference gene set. This set is used as a template, and the toxicity of an unknown compound is categorized based on the similarity or distance between the expression profile of the template and that of the compound.

These approaches have limitations for ab initio prediction of compound toxicity. First, they do not permit quantitative prediction of toxicity, which is essential to compare the toxicities among different compounds at a given dose. Second, the accuracy and generalizability of their toxicity predictions have not been fully examined. For the clustering approaches, it is methodologically not possible to validate the prediction result—without testing the accuracy and generalizability, the reference gene set for a certain type of toxicity may reflect a specific pharmacological response to the selected reference compounds. Third, the majority of studies using these approaches have not associated the expression profile with the predicted toxicity response in individual animals. It has been observed repeatedly and is accepted widely that the response could vary significantly because of genetic, environmental and physiological differences among animals.

Recently, more sophisticated and powerful machine learning algorithms have been applied to transcriptional profiling analysis. For example, a modified Fisher classification approach has been applied to distinguish between patients who do and do not have a good prognosis [15, 16]. A similar study used an artificial neural network [7]. However, in these studies, the variables to be predicted (e.g., survival time) were well defined. By contrast, because of the complexity of liver injury, no single measurement can comprehensively represent the degree of liver damage induced by compounds.

In this study, we developed a systematic approach for the ab initio prediction of compound-induced hepatotoxicity based on a rat liver global expression compendium. To overcome the three limitations discussed above, we developed a hepatotoxicity score based on five widely accepted clinical chemistry indicators and used the score to measure the severity of compound-induced liver damage [1]. Taking advantage of the score's quantitative nature, we applied a machine learning approach for the quantitative prediction of compound-induced liver damage based on expression profiles alone. The sensitivity and specificity of the established model were estimated with an independent data set. The established gene set and associated model can be used to assess the potential hepatotoxicity of compounds without prior knowledge of their toxicity mechanisms.

2 Method and Results

2.1 Experimental Materials, Methods and Data

Animal studies. Sprague-Dawley rats were treated with 49 hepatotoxic compounds and 10 non-hepatotoxic compounds (i.e., compounds without any previously observed or reported liver toxicity) as described elsewhere [17]. After 3-day treatment, liver samples were collected for RNA extraction and independent histopathology examination. Experiments were performed according to the guidelines in the NIH Guide for the Care and Use of Laboratory Animals. Serum was collected before necropsy and analyzed by standard methods to obtain clinical chemistry measurements. Table S1 (see http://www.jsbi.org/journal/GIW06/GIW06F016Suppl1.html) summarizes the treatments and the clinical chemistry measurements for the 267 liver samples used in this study.
RNA amplification, labeling and hybridization. Individual liver samples were profiled on a 25k rat liver microarray [17]. Detailed procedures for RNA preparation, amplification, labeling and hybridization have been described previously [5, 16, 17]. In brief, total RNA samples were extracted after DNase treatment. Five micrograms of total RNA from each sample was amplified into cRNA by an in vitro transcription procedure with oligo-dT primer. cRNA was labeled with Cy3 or Cy5 dye (CyDye, Amersham Pharmacia Biotech). For each amplified RNA sample, hybridizations were done in duplicate with fluor reversals. The reference cRNA pool was formed by pooling equal amounts of cRNAs from vehicle-treated control samples. After hybridization, slides were washed and scanned using a confocal laser scanner (Agilent Technologies). Fluorescence intensities of the scanned images were quantified, normalized and balanced.

Rat liver gene response compendium. A compendium of transcriptional responses to the 59 compounds in the rat liver was built with the previously reported 25k rat microarray [17]. Of the 49 hepatotoxins, 20 were administrated in both low and high doses, while 29 were administered only in high doses. Of the 10 non-hepatotoxins, all were administered in both low and high doses. For this study, 267 transcriptional profiles were obtained from rat liver after a 3-day treatment.

Unsupervised 2-dimensional hierarchical clustering uncovered distinctive patterns among hepatotoxins and non-hepatotoxins (Figure 1). We selected 2536 genes or reporting ESTs that changed more than 12% with p<0.01 in at least 3 profiles using error model-driven statistics [5, 15, 17] and the mean log ratio. These genes were significantly regulated in the compendium. The clustering was used to examine the general expression pattern; to emphasize the importance of co-regulation in clustering rather than the amplitude of regulations, we used the correlation as a similarity metric. Genes highly regulated by hepatotoxins did not overlap with genes regulated by non-hepatotoxins. Similarity in clustering over the compound profile dimension further indicated the distinction between clusters of non-toxins and clusters of toxins (Figure 1B, C). While subtle differences were observed, expression...
patterns among individual rats that received the same compound, whether a non-hepatotoxin or hepatotoxin, were substantially consistent (Figure 1B, C). Driven by these dominant patterns, profiles from repeats were often observed within the same cluster (Figure 1A). The overall expression patterns of the liver compendium revealed adequate reproducibility and sensitivity of the transcriptional profiling data for further analyses.

2.2 Analytic Methods and Results

Hepatotoxicity measurement based on clinical chemistry. A single quantitative measurement is needed to describe the severity of liver damage when predicting drug hepatotoxicity. Traditionally, a variety of clinical chemistry measurements [1] and histopathological classifications [21] have been used. For example, alanine aminotransferase (ALT) and aspartate aminotransferase (AST) in plasma are used as indicators of hepatocellular injury. Direct and total bilirubin (Tbil) are used to monitor potential cholestasis. Histopathology provides qualitative evaluation of liver injury at a cellular level, and cellular changes (e.g., necrosis, hypertrophy, steatosis and cholestasis) are often observed in drug-induced liver injury. However, because liver injury is complex, its severity cannot be sufficiently described by any one of these indicators. Therefore, we developed a method to integrate these traditional measurements into a single quantitative measurement, the hepatotoxicity score, to reflect the degree of liver injury and the numerous aspects of cellular damage.

Clinical chemistry-based hepatotoxicity score. We developed a hepatotoxicity score to estimate the severity of liver damage based on five widely used clinical chemistry measurements: ALT, AST, Tbil, alkaline phosphatase (ALP) and cholesterol (Chol). First, we converted the absolute value of each measurement into the distance between the normal value and the actual measured value defined as:

\[ D_{ij} = \frac{x_{ij} - \mu_{i,0}}{\sigma_{i,0}} \]  

where \( D_{ij} \) is the converted distance of ith measurement associated with jth profile, \( i=1,2,...,5, \) \( j=1,...,N=267 \). The \( x_{ij} \) is the original value of ith measurement associated with jth profile, \( i=1,2,...,5, \) \( j=1,...,N=267 \). \( \mu_{i,0} \) is the average of the ith measurement in the control experiments. \( \sigma_{i,0} \) is the standard deviation of the ith clinical chemistry from the control group. Then we normalized the five converted clinical chemistry measurements with sigmoidal transformation as defined in Equation (2) and Equation (3):

\[ D'_{ij} = \frac{1 - e^{-\alpha_i}}{1 + e^{-\alpha_i}} \]  

where

\[ \alpha_i = \frac{D_{ij} - \mu_{D_i}}{c_i \sigma_{D_i}} \]  

Herein, \( \mu_{D_i} \) is the average of the ith converted measurement in all experiments. \( \sigma_{D_i} \) is the standard deviation of the ith converted measurement from all experiments. \( c_i \) determines the range of linear transformation for the ith converted measurement. The standardization in the first step minimized differences of sensitivity and specificity across individual assays for the five measurements. In the second step, those standardized measurements were then normalized by sigmoid transformation. Because some measurements, such as ALT and AST, were often one or two factors away from the baseline level when liver damage occurred, the sigmoid transformation was preferred to retain outliers (abnormal values indicating severe liver damage) without compressing those values close to the threshold level in treated groups. For ALT, the most sensitive indicator of liver cell damage, we mapped values within 3 standard deviations, instead of 1 standard deviation, of the average to the most linear region of the sigmoid. After standardization and normalization, those measurements with a different scale could eventually be meaningfully combined into a single score as defined:
An Approach for Ab Initio Estimation

Hepatotoxicity Score = \sum_{i=1}^{5} \beta_i \times D_{i,j} \tag{4}

where we determined the weight factor \( \beta_s \) ad hoc based on previous knowledge about the sensitivity and specificity of individual clinical chemistry measurements for liver damage \cite{21}. The relevance of the combined score with five individual clinical chemistry parameters is described in the legend of Figure 2.

**Determination of hepatotoxicity score threshold for liver injury.** To establish the optimal threshold for the hepatotoxicity score, we first determined the total number of potential liver injuries detected by the combination of the five clinical chemistry measurements. These detected injuries were used as the substitute for the “gold standard” of liver damage; specifically, any animal with one of four clinical chemistry measurements (AST, ALT, ALP and Tbil) at least two standard deviations away from the average level in the control group was considered to have liver damage. Plasma cholesterol level might be related to liver function, but we considered an abnormal cholesterol level indicative of liver damage only if it co-occurred with at least one additional abnormal clinical chemistry measurement. Although the specificity of individual clinical chemistry parameters for liver injury was often questionable, our strategy aimed to minimize any false-negatives.

Using the 44 liver injuries detected from the 267 liver samples in the current study and 9 liver samples in a separate study, we determined a threshold for the hepatotoxicity score for liver injury by minimizing the sum of false-positives and false-negatives (Figure S1(see http://www.jsbi.org/journal/GIW06/GIW06F016Suppl1.html)). With the threshold selected (-0.25), the sensitivity and specificity of the hepatotoxicity score was estimated to be 90.9% and 98.7%, respectively (Table 1). In contrast, individual clinical chemistry parameters had much lower sensitivity for identifying the 44 liver injuries—56.8% with AST, 79.5% with ALT, 36.4% with ALP, 6.8% with Tbil and 25% with Chol (Figure 2).

<table>
<thead>
<tr>
<th>Table 1: Comparison of hepatotoxicity score-predicted liver injuries vs. expected liver injuries.</th>
<th>Table 2: Comparison of expression model-predicted liver injuries vs. expected liver injuries in the training data set.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Liver Injuries</td>
<td>Expected Liver Injuries</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Yes</td>
<td>40</td>
</tr>
<tr>
<td>No</td>
<td>4</td>
</tr>
</tbody>
</table>

To further evaluate the score’s accuracy, we assessed histopathological findings for the false-negatives and false-positives obtained by the hepatotoxicity score. Among the four false-negatives, one was metformin, a non-hepatotoxin, at 900 mg/kg/day; two were low-dose (subtoxic) hepatotoxins, TNF-alpha at 0.01 mg/kg/day and tamoxifen at 5 mg/kg/day; and one was a high-dose hepatotoxin, monocrotaline at 50 mg/kg/day. None of the three hepatotoxin-treated samples had histologic abnormalities, suggesting that the hepatotoxicity score had higher specificity than the individual clinical chemistry measurements. All of the three false-positives occurred in high-dose hepatotoxin groups. Specifically, the positives detected by the hepatotoxicity score, but not by any of the clinical chemistry measurements, occurred in the estradiol glucuronide (10 mg/kg/day) and aspirin (150...
Figure 2: Developing a hepatotoxicity score based on five clinical chemistry (CC) measurements: Tbil, ALP, ALT, AST and cholesterol (Chol). The mean of the individual CCs in the control group is indicated by the blue dotted line. The threshold of individual CCs and the hepatotoxicity score for liver injury are shown by the red dotted line. Liver injury as revealed by individual CCs is indicated by red dots. For the 44 liver injuries detected, levels of ALT, AST, ALP and Tbil were abnormal in 35, 25, 16 and 3, respectively; cholesterol level was abnormal in 11. The hepatotoxicity score is the weighted sum of normalized individual CC scores after standardization. The weight for each normalized CC score was adjusted according to the specificity for liver injury. In particular, the weights for ALT and AST were 3 and 1.5, higher than 1, the assigned weight for Tbil. Because ALP has low specificity, its weight was assigned as 0.5. Cholesterol was counted only when any one of the four other CCs was also abnormal. The hepatotoxicity score detected 40 cases of liver injury (green circles).

mg/kg/day) groups. Cellular abnormality was observed in estradiol-treated, but not aspirin-treated animals. Among the three estradiol glucuronide-treated rats, one showed liver injury by ALT and ALP as well. The evidence from histopathology examination and clinical chemistry measurements for other rats in these groups suggested that these false-positives might be due to detection of mild liver injury by the hepatotoxicity score, but not by any single clinical chemistry measurement.

Prediction of compound-induced liver damage based on transcriptional profiles by a multilayer perceptron neural network classifier. To build a classifier for predicting hepatotoxicity based on transcriptional profiles, we randomly divided the compendium of 267 profiles with their associated clinical chemistry measurements into a training data set (212 profiles, 80%) and an independent validation data set (54 profiles, 20%). Our procedure for ab initio prediction of hepatotoxicity based on transcriptional profiles is illustrated in Figure 3.

Briefly, we selected the 238 genes significantly differentiated between the non-hepatotoxin and hepatotoxin groups as candidate marker genes for the classifier based on their error-weighted log ratio (p value <1e-5). Accession numbers and annotations for these genes are given in Table S2(see http://www.jsbi.org/journal/GIW06/GIW06F016Suppl.html); raw data for these genes are available on
An Approach for Ab Initio Estimation

request.

Figure 3: Diagram for *ab initio* prediction of hepatotoxicity based on transcriptional profiles.

The error-weighted log ratio is defined as

\[ X_{devi} = \frac{\log x_i}{\sigma \log x_i} \]  \hspace{1cm} (5)

where \( X_{devi} \) is the error \((\sigma x_i)\)-weighted log ratio \((x_i)\) of the \(i\)th gene. The false discovery rate associated with the selected \(p\) value has been assessed by Monte Carlo simulation.

To reduce the number of dimensions, we transformed the \(X\)dev for each significantly regulated gene by wavelet transformation with Daubechie’s wavelet function at a level of 5. This transformation retained the main regulation information across treatment groups for each gene, but reduced the variable dimension to 31. This reduction in turn reduced the chance that the model obtained from the training data set would be overfitted. The classifier for hepatotoxicity score prediction was then trained based on the transformed features and the associated clinical chemistry-based hepatotoxicity score in the training set.

Establishment of two-layer perceptron neural network classifier. The first layer of the network is defined as

\[ a_j^{(1)} = \sum_{i=1}^{d} w_{ji}^{(1)} x_i + b_j^{(1)} \]  \hspace{1cm} (6)

Where \(i = 1, \ldots, 31; j = 1, \ldots, M\), \(x_i\) represents \(i\)th input feature from transformed expression profiles; \(a_j^{(1)}\) represents \(j\)th variables in the first layer of the network. Here \(w_{ji}^{(1)}\) are the elements of the first layer weight matrix and \(b_j^{(1)}\) are the bias parameters associated with the hidden units.

The output of the network is defined as \(y_k = a_k^{(2)}\) Equation (7) where \(k = 1 \ldots n\) and

\[ a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + b_k^{(2)} \] where \(z_j = \tanh(a_j^{(1)})\) Equations (8, 9).
We established the best classifier by cross-validation in the training set [6, 11].

**Evaluation of the hepatotoxicity score-based classifier.** We examined the specificity and sensitivity of the trained model in the training data set using the pre-established hepatotoxicity score threshold. Compared with the previously identified liver injuries in the training set, 12 false-positives and 10 false-negatives were obtained with the trained prediction model. The sensitivity and specificity of the trained model were 70% and 93%, respectively, and its overall prediction accuracy was 89.6%. Histopathologic examination showed liver damage among 5 of the 12 false-positive samples despite normal clinical chemistry measurements, suggesting that the trained model based on expression profiles has the potential to predict liver injuries detected by pathological examination alone (Table 2).

Figure 4: Sensitivity and specificity of the built model estimated with the independent validation data set. The regulation of 238 reported genes (left panel) reveals a major pattern associated with liver damage in the validation data set (heatmap scale: -1~1). The hepatotoxicity score (blue dots) and the predicted score (blue line) are shown in the middle panel. The threshold for liver damage is indicated by the red dashed line. The expected positive (EP) based on individual clinical chemistry measurements, and predicted positive (PP) based on hepatotoxicity scores are indicated by black shading in the right panel. The false-positive (FP) and false-negative (FN) predictions for liver damage, determined by comparing PP with EP, are also indicated by black shading in the right panel. A high hepatotoxicity score predicted from the profiles (indicated by black labels in bold) may result from different contributions of individual genes to the hepatotoxicity score in the trained model.

To determine the expected error and generalizability of the trained model for hepatotoxicity prediction, we evaluated the trained model in the validation data set, which had not been used previously.
in establishing the model (Figure 4). The predictions from the built classifier based on this data set were compared with the previously identified liver injuries. Among the 54 predictions, 5 false-positives and 1 false-negative were obtained. Using the pre-established hepatotoxicity score threshold, the sensitivity, specificity and accuracy of the trained model in the validation data set were 90.9%, 88.4% and 88.9%, respectively. The similarity of the sensitivity and specificity to those in the training set support the generalizability of the built model in predicting compound-induced liver damage solely based on expression profiles.

3 Discussion

We developed a model for ab initio prediction of hepatotoxicity based on 238 marker genes using 212 transcriptional profiles and validated it using 54 independent transcriptional profiles. The comparable accuracy between the training and validation data sets indicates acceptable generalizability of the trained model, suggesting a reasonable coverage of the toxicities and an adequate number of samples in the training data set. Hence, it might be possible to use the gene set and the associated model derived from our method to assess compound-induced liver damage within the specified range of accuracy.

In contrast to the majority of reported approaches in toxicogenomics [3, 4, 9, 12], we estimated liver toxicity by a single quantitative measurement, the hepatotoxicity score, which incorporated five clinical chemistry measurements. The approach of integrating multiple measurements into one continuous index for liver toxicity was critical for predicting the degree of liver toxicity based on transcriptional profiles. Examination of the false-positives and -negatives from our built model suggested that expression profiling-based liver injury prediction could be more sensitive and specific than the combination of clinical chemistry parameters. For example, the five false-positives detected in the validation data set included profiles from one rat each in the dimethylformamide (1000 mg/kg/day) group, the tetracycline (500 mg/kg/day) group, the diethylnitrosamine (100 mg/kg/day) group, the L-ethionine (50 mg/kg/day) group and the levofloxacin (200 mg/kg/day) group (Table 3). Histopathological changes were observed in the dimethylformamide-, tetracycline- and L-ethionine groups; the diethylnitrosamine group was not examined, although histopathological changes have been reported at this dose [18].

Table 3: Summary of histopathology for all false predictions in the training and validation data sets.
Of note, 83% of compounds studied were selected based on their reported hepatotoxicity in the literature, yet only 16.5% of rats treated had chemistry or histopathologic evidence of liver injury. The absence of liver toxicity in the majority of the rats that received hepatotoxins for 3 days reinforces the close relationship between the level or duration of exposure and the degree of toxicity. Because our classifier is built based on the expression profiles from short-term exposure to these compounds and their potential toxic effects in the liver, our study cannot determine the extent to which these effects would occur with long-term exposure and would elicit substantial toxicity. Hence, our marker gene prediction model focuses on providing a measure of induced liver injury based on expression profile rather than on characterizing a compound as a hepatotoxin or a non-hepatotoxin. Additional, long-term study may help to further validate our model for assessing compound-induced liver damage at higher exposure levels. Classification of hepatotoxins beyond the level of exposure would require a better knowledge of mechanisms underlying their toxicity.

The analytical approach we developed for predicting hepatotoxicity based on transcriptional profiles has important utility and implications in both drug discovery and basic biological research. In toxicogenomics, this is the first method that enables ab initio prediction of drug toxicity in a quantitative manner. Without prior pharmacological or toxicological knowledge, this approach estimates potential liver toxicity based solely on transcription profiles. Increasing the size and diversity of our compendium could further improve the model’s predictive sensitivity and specificity.

Our approach for hepatotoxicity prediction based on transcriptional profiles can be improved on. For example, a variety of different strategies could be implemented for selecting and optimizing the hepatotoxicity score, marker genes and transformation of original data. Using systematically optimized weight factors for the hepatotoxicity score (instead of ad hoc determined weight factors) or additional strategies (pathological readouts and leave-one-out cross-validation, instead of 1-out-of-5-fold validation) in selecting the best classifier may improve the performance of our classifier. Furthermore, an exhaustive optimization of the marker genes and associated transformation may improve the accuracy of the prediction model. Although we used an artificial neural network algorithm, other supervised machine learning algorithms, such as Bayesian network and the supporting vector machine, can also be applied. Nevertheless, herein we demonstrate that it is feasible to quantitatively assess compound-induced liver toxicity based on transcriptional profiles.

Acknowledgments

We thank Drs. Thomas Skopek and Frank Sistare for their helpful comments and advice, and the Rosetta Gene Expression Laboratory for generating the highly reproducible expression profiling data. The authors are employees of Merck & Co. Inc. (XD, YDH, HD, PYL, CJR, RGU) or Abbott Laboratories (JFW).

References


An Approach for Ab Initio Estimation


