Prediction of Drug Induced Liver Injury Using Molecular and Biological Descriptors

Christophe Muller¹, Dumrongsaik Pekthong²,³, Eliane Alexandre², Gilles Marcou¹, Dragos Horvath¹, Lysiane Richert*²,⁴ and Alexandre Varnek*¹

¹Laboratoire de Chemoinformatique, UMR 7140 CNRS, Université de Strasbourg, Strasbourg, 67000, France
²KaLy-Cell, 20A rue du Général Leclerc, 67115 Plobsheim, France
³Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Naresuan University, Phitsanulok, 65000, Thailand
⁴Université de Franche-Comté, Besançon, France

Abstract: In this paper we report quantitative structure-activity models linking in vivo Drug-Induced Liver Injury (DILI) of organic molecules with some parameters both measured experimentally in vitro and calculated theoretically from the molecular structure. At the first step, a small database containing information of DILI in humans was created and annotated by experimentally observed information concerning hepatotoxic effects. Thus, for each compound a binary annotation “yes/no” was applied to DILI and seven endpoints causing different liver pathologies in humans: Cholestasis (CH), Oxidative Stress (OS), Mitochondrial injury (MT), Cirrhosis and Steatosis (CS), Hepatitis (HS), Hepatocellular (HC), and Reactive Metabolite (RM). Different machine-learning methods were used to build classification models linking DILI with molecular structure: Support Vector Machines, Artificial Neural Networks and Random Forests. Three types of models were developed: (i) involving molecular descriptors calculated directly from chemical structure, (ii) involving selected endpoints as “biological” descriptors, and (iii) involving both types of descriptors. It has been found that the models based solely on molecular descriptors have much weaker prediction performance than those involving in vivo measured endpoints. Taking into account difficulties in obtaining of in vivo data, at the validation stage we used instead five endpoints (CH, CS, HC, MT and OS) measured in vitro in human hepatocyte cultures. The models involving either some of experimental in vitro endpoints or their combination with theoretically calculated ones correctly predict DILI for 9 out of 10 reference compounds of the external test set. This opens an interesting perspective to use for DILI predictions a combination of theoretically calculated parameters and measured in vitro biological data.

Keywords: Biological descriptor, drug-induced liver injury, human hepatocyte cultures, machine-learning methods, molecular descriptors.

INTRODUCTION

Drug-Induced Liver Injury (DILI) is the leading cause of acute liver failure and the most common adverse event causing drug non-approval and drug withdrawal. Approximately 40% of new drug candidates fail in clinical trials because of serious toxic events that remained unrecognized in preclinical studies. While 85% of cardiovascular, 88% of gastrointestinal and 90% of hematomal toxicity can be assessed in animal toxicity tests, predictions of hepatotoxicity for humans still suffer from less than a 50% success rate [1] which is inappropriate for drug screening campaigns. Besides animal model, the state of the art for hepatotoxicity early assessment is based on in vitro studies of human hepatocytes [2]. Some efforts have been made to develop computational tools for DILI predictions based on mechanistic considerations. Thus, Bhattacharya et al. [3] reported predictive physiological model (DILIsym™) to understand drug-induced liver injury which focuses on reactive metabolite-induced DILI in response to administration of acetaminophen. Yet, there is a clear need for fast, reliable and easily accessible computational tools.

Generally, Quantitative Structure-Activity Relationships (QSAR) are widely used in toxicity predictions [4]. However, only few theoretical studies devoted to hepatotoxicity predictions have been reported. Thus, one may mention QSAR models based on one- and two-dimensional descriptors [5], on molecular field analysis [6], or on identification of toxicophores [7]. Their predictive performance is still lower than that of models obtained for other toxicities.
Recently, new types of models using information of biological in vitro or in vivo data as descriptors were reported [5b, 8]. It has been shown that they predict hepatotoxicity much better than conventional QSARs based on molecular descriptors issued from chemical structure. Thus, Low et al. [8b] in QSAR modeling of drug-induced hepatotoxicity in the rat combined molecular descriptors and results in vivo tests considered as “biological” descriptors. In another study, Liu et al. [8a] related 13 hepatotoxic effects in humans to classify drugs as DILI or non-DILI: if one side effect out of 13 was observed, the drug was considered DILI. Results reported in references [5b, 8a] clearly demonstrate high predictive performance of the models involving biological descriptors.

The aim of this work was to obtain a model able to predict hepatotoxicity in humans. At the first step, using a text mining technique we have built a database containing 424 reference compounds for which hepatotoxic effect were reported in the literature. In this database, each compound was annotated by 7 additional endpoints known as DILI markers: “Cholestasis”, “Oxidative Stress”, “Mitochondrial injury”, “Cirrhosis and Steatosis”, “Hepatitis”, “Hepatocellular” (i.e. apoptosis/necrosis), and “Reactive Metabolite”. Then, classification models “DILI/non-DILI” were built using various machine-learning methods and different types of descriptors: (i) molecular descriptors generated directly from molecular structures and (ii) experimental values of measured in vivo 7 endpoints represented by binary numbers. An effort also was made to obtain classification models for the above 7 endpoints using solely molecular descriptors. Finally, the models based on the combination of experimental and predicted endpoints were obtained. Calculations performed on the independent test sets showed that the models based on solely molecular descriptors do not perform well, whereas those involving experimental endpoints possess high predictive ability. Practical application of these models to new molecules is, however, very limited because of the difficulties to assess in vivo endpoints experimentally. For this reason, we investigated the use of in vitro data as “biological” endpoints and applied the modified models to a dataset of 10 reference drug molecules (9 DILI and 1 non-DILI) for which required endpoints were assessed in human hepatocytes.

**MATERIALS AND METHODS**

**Computational Procedure**

Data Preparation. A list of molecules relating to drug induced liver injury (DILI) was extracted from Toxnet [9] which is a website gathering databases on toxicology, hazardous chemicals, environmental health, and toxic releases. Toxnet constitutes the principal resource of information of the U.S. National Library of Medicine. At the first step, the search has been performed using different DILI related terms (e.g., “DILI”, “Drug-Induced Liver Injury” or “Liver Failure”) as keywords. Similar procedure was applied to search for seven additional endpoints known as DILI markers: “Cholestasis”, “Oxidative Stress”, “Mitochondrial injury”, “Cirrhosis and Steatosis”, “Hepatitis”, “Hepatocellular” (i.e. apoptosis/necrosis), and “Reactive Metabolite”. Each query specified that the biological effect had to be observed in humans. If a molecule matched, at least, one of the searched terms, its CAS (Chemical Abstract Service) number was extracted and a label “1” was assigned in the corresponding field. If the molecule was known for its potential therapeutic role in the treatment of the mentioned liver effect, a label “0” was assigned to this molecule in the corresponding field.

The Toxnet search resulted in a dataset of 546 compounds labeled “0” or “1” for each endpoint. A second dataset composed of 217 drugs DILI and 147 non-DILI related to the above endpoints was extracted from the literature. These two datasets were merged in the database of 805 compounds including individual molecules, natural extracts and mixtures of which 493 drug molecules were kept for the modeling. For each molecule in this set, information about its DILI activity (“DILI” or “non-DILI”) and, at least, about one from the above mention endpoints, was available.

Data Curation. Using retrieved CAS numbers, SMILES (Simplified Molecular Input Line Entry specification) codes of selected drugs were extracted from the “PubChem compound” database [10] and then their 2D structures were automatically generated with the InstantJChem/ChemAxon software [11]. Data curation procedure has been performed following the recommendations by Fourches et al. [12]. Accordingly, inorganic compounds were removed from dataset using InstantJChem. Then structures were standardized with Standardizer from ChemAxon following the routine: aromatize (basic), Remove Fragment (keep largest fragment). In addition, visual inspection was performed in order to check if active compound (rather than other mixture components) was kept. If the active compound was not clearly identified then the whole mixture was discarded. Automatic search of duplicates was performed using the EdiSDF software (freely available at http://infochim.u-strasbg.fr/). For duplicates having endpoints with similar labels, only one of them was kept. If labels of endpoints differed, expert opinion was needed to choose an entry to be kept. Finally, for each selected molecule the expert using original bibliographical sources verified the annotation of DILI and all seven endpoints.

The curated dataset consists of 424 drugs including 247 DILI and non-DILI compounds. With respect to end-points, the dataset contains 168 drugs inducing cholestasis, 61 cirrhosis or steatosis, 76 hepatitis, 58 mitochondrial injury, 41 oxidative stress, 166 hepatocellular (necrosis or apoptosis), and 67 reactive metabolites, see data distribution on Fig. (1).

Descriptors. ISIDA fragment descriptors have been generated by the ISIDA/Fragmentor program [13] (freely available at http://infochim.u-strasbg.fr/). Two classes of the fragments were used: “sequences” (I) and “augmented atoms” (II). For each class, three sub-types were defined: AB, A and B. Sequences represent the shortest paths from one atom to another one: (AB) represented sequences of atoms and bonds, (A) of atoms only, and (B) of bonds only. Sequences could also be represented as Atoms Pairs (AP) where only terminal atoms and topological distance between them were represented explicitly. An “augmented atom” represents a selected atom with its environment including
sequences of AP, AB, A and B types issued from this atom. Descriptors of length from 2 to 10 atoms were calculated for sequences and from 2 to 6 atoms for augmented atoms. Variation of fragments’ types and size resulted in generation of 300 initial pools of descriptors.

Machine learning methods. Different machine learning methods have been used for structure-activity modeling of both DILI and different end-points: Random Forest, SVM, ASNN and AMORE.

The Random Forest (RF) [14] method is composed of a user defined number of trees (100 trees in this study). Each tree is constructed on a randomly selected subset of descriptors from initial pool of descriptors. The overall prediction is made by a majority vote of all trees. In this study 10 randomly selected descriptors were used to construct each tree. The WEKA implementation of RF was used.

Support Vector Machine (SVM) is a popular classifier, which searches for the hyperplane that separates two classes of instances in the feature space. Construction of models was performed using the LIBSVM software [15] with Tanimoto coefficient as a kernel. The cost parameter was systematically varied from 1 to 101 incrementing by 10. Its value providing with the best averaged balanced accuracy in 5-fold external cross-validation has been selected. The instances were weighted according to the proportion of the class they belong to.

AMORE is a module in R package to construct neural networks [16]. A neural network (NN) mimics behavior of the human brain where numerous neurons are interconnected. Typically, an artificial NN have 3-layers architecture including input, hidden and output layers. Each neuron in the input layer corresponds to one descriptor, whereas in the output layer it corresponds to one modeled activity. Each neuron of the hidden layer receives weighted signals from the input layer and then, as a function of received signals, sends a signal to each neuron of the output layer. Error back-propagation algorithm has been used to train the models. Early stopping procedure [17] was used to avoid over-fitting.

Associative neural networks (ASNN) [18] approach dealt with a combination of several neural networks and the k-nearest neighbor technique. The ASNN program uses the same early stopping procedure as AMORE. Two modeling strategies were used in AMORE and ASNN: single task learning (only one property is modeled) and for multi task learning (many properties are modeled in parallel). It has been shown that MTL can significantly improve the models’ performance if simultaneously trained properties are somehow related [19].

Model validation. Predictive performance of the models was assessed in 5-fold cross-validation. In this procedure, the initial data set is randomly split into n=5 equal portions. This allows one to form 5 non-overlapping training and test sets, each containing, respectively, 4/5 and 1/5 molecules of the initial set. The models are built on each training set followed by its validation on the corresponding test set. These calculations are performed for all 5 folds; thus, each molecule is predicted once. The workflow shown on Fig. (2) show that parameterization of the machine-learning algorithm and the selection of a pool of molecular descriptors are performed using only the training (or modeling) sets. Thus results reported for the test sets are rigorously external because the models are applied to the molecules never seen by the program during model building.

In this study, Balanced Accuracy (BA) parameter [20] was used to estimate quantitatively the models’ performance.

\[
BA = \frac{1}{2} \left( \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right)
\]

In this equation, the information about true positive (TP), false positive (FP), true negative (TN) and false negative (FN) examples is used. For each data set, BA was calculated of the average of values calculated for each fold of cross-validation. Notice that for binary classification problem, BA = 0.5 corresponds to random attribution of classes, and BA = 1 – to ideally predictive model.

In some cases, due to pure statistical chance some descriptors may correlate with the modeled activity without any real meaning in capturing structure-activity relation. In order to check if the obtained models could be explained by such “chance correlations”, the Y-scrambling [21] method can be particularly useful. In Y-scrambling, the activity data are randomly shuffled to change their true order. This alter correct attribution of activities to a given descriptors’ vector and, therefore, should destroy any meaningful structure-activity relation. Thus, the performance of the model fitted for randomly reordered activity values should be much lower than that the model obtained for the actual activity values. In this study, for each machine-learning method 30 scrambling steps have been performed for each selected descriptor pools followed by calculations of corresponding BA values. The upper bound of a confidence interval of 95% calculated for each BA distribution (BA_{scr}) has been compared with the Balance Accuracy of the model (BA_{model}) built on the initial data. The model was kept if \( \text{DBA} = | \text{BA}_{\text{model}} - \text{BA}_{\text{scr}} | \) was larger than 0.01, otherwise it has been rejected.

IN VITRO ASSESSMENT OF HEPATOTOXICITY MARKERS

Reagents and Materials

Human hepatocytes were isolated as previously described [22] by a two-step collagenase perfusion technique and
cryopreserved/thawed as described by Alexandre et al. [23].

Media, fetal calf serum, and supplements for culture were from Life Technologies (Fisher Scientific, Illkirch Graffenstaden, France). DMSO and analytical grade solvent were purchased from Sigma. All other chemicals (diclofenac, valproic acid, chlorpromazine, cyclosporine A, acetaminophen, ticlopidine and warfarin) were provided by Sigma-Aldrich (Saint Quentin Fallavier, France); troglitazone and rosiglitazone were from Cayman chemical company (Euromedex, Vendenheim France). Single use laboratory plastic ware, PP-centrifuge tubes and 96-well round-bottom plates were obtained from Nunc (Dominique Dutscher, Brumath, France).

**Human Hepatocyte Treatments**

Cryopreserved human hepatocytes were either maintained in suspension under continuous shaking up to 3 hours that has been shown to allow high viability [24] or plated [25] or cultured as monolayers. Cells were exposed to a set of ten reference compounds: acetaminophen, amiodarone, chlorpromazine, cyclosporine A, diclofenac, rosiglitazone, ticlopidine, troglitazone, valproic acid, and warfarin.

A series of measurements were performed in order to obtain in vitro values of the endpoints used in the modeling. Hepatocellular necrosis was assessed by a cytotoxicity assay, the Cell titer Blue® test [26]. Cholestasis was assessed according to Chatterjee et al. [27]. Oxidative Stress and Mitochondrial injury were assessed according to Li [28] with some modifications. Steatosis was assessed according to Borenfreund and Puerner [29] with some modifications.

**RESULTS AND DISCUSSION**

**Modeling of Individual Endpoints Using Solely Molecular Descriptors**

In order to develop QSAR models of DILI and 7 endpoints, 300 subsets of ISIDA substructural fragments descriptors were generated and used as initial descriptors pools in combination with six machine-learning methods: SVM, RF, ASNN (STL), ASNN (MTL), AMORE (STL), and AMORE (MTL). Cross-validation results, based on the modeling sets (datasets that exclude the external test set compounds) for the optimum parameters of each machine-learning algorithms and the optimum subset of ISIDA substructural fragments descriptors are summarized in Fig. (3). At the second step, Y-scrambling experiments have been performed and compared to the unscrambled situation; the results are summarized on Fig. (4). Only few models survived in the scrambling tests. They were applied to the external test set of the main cross-validation loop and the results are reported in the Fig. (5). Thus, no acceptable model was obtained for the OS and HS endpoints. For other 6 endpoints, performance of the best models varies in the range BA = 0.58 - 0.70. The SVM models perform better for all endpoints except of CS for which AMORE (MTL) model was found the best.

Predictive performance of SVM and AMORE (MTL) models for DILI is rather weak (BA = 0.66), which is similar to previously reported results for QSAR models of hepatotoxicity in human based on molecular descriptors [5b, 8]. This shows that the DILI phenomenon integrating different mechanisms of action of toxic compounds is too complicated to be described by conventional QSAR models based on this set of molecular descriptors solely.

**DILI modeling using in silico predicted endpoints.** Since obtaining in vivo endpoints is very expensive and time consuming, we made an effort to build a model based on predicted endpoints. Thus, the model has been built on CH, HC, MT and RM endpoints predicted with SVM and CS calculated with AMORE (MTL) methods. HS and OS were not used because no acceptable models were obtained for these endpoints. The model’s performance assessed in 5-CV (BA=0.54) is even worse than that of the QSAR model for DILI based on molecular descriptors (BA=0.66). This drop of performance might be explained by a noise issued from additional modeling steps. Since most of in silico predicted
endpoints don’t reach a BA of 0.6, one can hardly expect high performance of the models involving these endpoints.

**DILI modeling using in vivo observations.** Relatively weak performance of QSAR models based on molecular descriptors motivated us to change the paradigm and to use *in vivo* observed endpoints as binary variables: “1” if the drug induces the endpoint, “0” otherwise. The AMORE program, well suited to work with binary variables, was used to build the model (*Model 1*). The model’s performance assessed in 5-folds cross-validation is rather high (BA = 0.89). The model successfully passes the Y-scrambling test: \( BA_{scr} = 0.59 \) which is much lower than that obtained for the model.

**DILI predictions using in vitro observations.** The above results show that the models based on *in vivo* endpoints perform much better than those involving molecular descriptors. However, application of these models to new molecules is limited because of the difficulties to assess *in vivo* endpoints experimentally. The question arises whether one can replace *in vivo* with *in vitro* data upon the models’ application to external dataset.

To investigate this, a set of 10 drug molecules has been selected: diclofenac, troglitazone, valproic acid, chlorpromazine, amiodarone, cyclosporine A, acetaminophen, ticlopidine, warfarin and rosiglitazone (9 DILI and 1 non-DILI). Unfortunately, we were not able to apply *Model 1* based on 7 endpoints, because only 5 endpoints - CH, CS, HC, MT and OS - can be measured experimentally by Kaly-Cell in human hepatocytes.

Therefore, the model based on 5 *in vivo* endpoints (*Model 2*) was built with AMORE on the entire dataset from which 10 drug molecules of the test set were excluded (totally 414 molecules). Another model (*Model 3*) was obtained using five *in vivo* endpoints (CH, CS, HC, MT and OS) and 1 complementary endpoint (RM) survived in Y-scrambling tests. Both models display similar predictive performances in 5-CV (BA = 0.86 and 0.87 for Models 2 and 3, respectively) and survive in Y-scrambling tests (\( BA_{scr} = 0.56 \) and 0.55, respectively).
These models were applied to 10 molecules for which the above 5 endpoints were measured in vitro in human hepatocytes. Despite disagreement between in vivo/in vitro labels for some endpoints, both models correctly predict DILI for 9 out of 10 compounds. Thus, in most of cases, ensemble of in vitro data leads to correct DILI predictions using Models 2 and 3, even if for some particular endpoints in vivo and in vitro labels differ. The only incorrectly predicted compound is warfarin classified experimentally as DILI. However, in FDA database this compound is labeled as “less-DILI-concern”. Thus, the implication of this drug in liver injury is not firmly established and, in this context, our “incorrect” prediction may make sense.

Comparison with previously reported models. Although direct comparison of the performance of our models with previously reported ones is difficult because they were trained and validated on different data sets, some rough analysis can be done.

Our work confirms previously reported observations concerning very weak performance of the models involving solely molecular descriptors. Thus, according to our results, BA varies in the ranges 0.61 - 0.65 (see Fig. 5) similarly to BA =0.55 – 0.61 reported by Low et al. [8] On the other hand, involvement in vivo experimental data as “biological descriptors” significantly improves the quality of predictions.

Liu et al. [8a] successfully applied 13 hepatotoxic side effects to classify drugs as DILI or non DILI. The BA values (calculated from the data in table 1 in reference [8a]) for two external test sets are rather high (0.91 and 0.76, respectively). The models by Low et al. [8b] involving in vivo liver toxicogenomics data (transcripts) resulted in BA = 0.72 - 0.78. Qualitatively, this corresponds to high performance of our models trained on 8 in vivo endpoints (BA= 0.87-0.89).

Application of predicted instead of experimental endpoints in the modeling significantly decreases the accuracy of predictions. Thus, BA = 0.54 in 5-CV obtained in this work is less good than the results reported by Liu et al. [8a] for two largest test sets: BA = 0.62- 0.66, calculated from the data in table 2 in [8a].

Using in vitro data upon application of our models initially trained on in vivo data appeared to be a reasonable and efficient strategy of DILI predictions, since BA values were respectively 0.86 for Model 2 and 0.88 for Model 3. Our calculations correctly predicted human in vivo DILI for 9 out of 10 compounds, demonstrating the performance of the model based on combined “fragment” and “in vitro biological” data and opens an interesting perspective for cost-reduction and in vivo predictivity of DILI in drug design campaigns.

CONCLUSION

In this paper, we report quantitative structure-activity models linking in vivo Drug-Induced Liver Injury (DILI) of organic molecules with in vivo observed endpoints (Cholestasis, Oxidative Stress, Mitochondrial injury, Cirrhosis and Steatosis, Hepatitis, Hepatocellular apoptosis/necrosis, and Reactive Metabolite) as well as some descriptors derived directly from molecular structure. These models display good prediction performance (Balanced Accuracy is about 0.9 in 5-folds cross-validation). On the other hand, models involving molecular descriptors solely poorly perform in DILI and other proposed endpoints predictions. Since obtaining of in vivo endpoints for new compounds is time consuming and expensive task, we suggested applying in the models their in vitro analogues measured in KalyCell company. The model involving in vitro instead of in vivo endpoints has been successfully applied to a dataset of 10 drug molecules: DILI labels were correctly predicted for 9 of them. These results show that although the labels for particular endpoints in vitro and in vivo may differ, the common application of endpoints as descriptors may lead to correct DILI predictions. A modeling of in vivo DILI based on in vitro measured endpoints would be desirable, but still is hardly possible because of data availability problem.

ABBREVIATIONS

| BA   | Balanced Accuracy |
| CH   | Cholestasis       |
| CS   | Cirrhosis and Steatosis |
| DILI | Drug-Induced Liver Injury |
| HC   | Hepatocellular (necrosis/apoptosis) |
| HS   | Hepatitis         |
| MT   | Mitochondrial Injury |
| MTL  | Multi Task Learning |

n-CV = n-folds cross-validation

Fig. (5). Balanced Accuracies averaged over 5 external sets for the of endpoints using models surviving in Y-scrambling test.
Prediction of Drug Induced Liver Injury

NN = Neural Network
OS = Oxidative Stress
RF = Random Forest
RM = Reactive Metabolite
STL = Single Task Learning
SVM = Support Vector Machine

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher’s web site along with the published article.

Supplementary material contains (i) the list of chemicals used in modelbuilding, (ii) the list of key words used to mine the literature, (iii) statistical parameters of the models obtained in this work and those reported in the literature, and, (iv) some information about Tanimoto kernel based on ISIDA descriptors.

REFERENCES


