Carolina Center for Computational Toxicology

Funded by U.S. EPA Cooperative Agreement STAR RD 833825

Progress Report for Project Period 04/01/09 - 03/31/10 (Year 2)

Report Date: 04/21/10

Ivan Rusyn, Principal Investigator

Project 3:
Development of validated and predictive Quantitative Structure-Toxicity Relationship models that employ both chemical and biological descriptors of molecular structures and take into account genetic diversity between individuals

Preliminary Data and Work Progress

This project is focused on the development of novel computational approaches to link chemical structure, *in vitro* data, and potential adverse health effects. Our ultimate goal is to improve the prediction of chemical toxicity (both in human and in animals) using both the explicit information about chemical structure and emerging data generated in short-term, cell based (*in vitro*) assays. In addition to advancing the traditional QSAR modeling approaches as applied to toxicity prediction using ToxCast and other data, we have continued to develop novel Quantitative Structure *In vitro-In vivo* Relationship (QSIIR) workflows to enable robust and predictive cheminformatic models of animal toxicity. To this end, we have developed several innovative computational approaches that enable: (i) in vitro-in vivo toxicity database compilation and curation; (ii) creation of "hybrid" (i.e., using both chemical and biological descriptors) compound attributes and development of corresponding hybrid cheminformatic models; (iii) model validation and their application as virtual filters for the evaluation of chemical toxicity of untested chemicals in external chemical libraries. Most studies have been done using an extensive collection of data enabled by the ToxCast™ project. The resulting computational toxicity predictors are expected to directly evaluate the liver toxicity potential of chemicals and therefore, prioritize them for future testing as well as help discard chemicals that are likely to have hazardous effects.

Results to Date

A large part of our research effort has been devoted to the analysis of ToxCast Phase I data in collaboration with projects 1 and 2. First, conventional, i.e., chemical descriptors-based QSAR approach was applied to data in the toxicity reference database (ToxRefDB). The purpose of this analysis was (i) to identify endpoints with sufficient amount of data to enable model development in principle; and (ii) to set a reference point with respect to the predictive power of conventional models for subsequent comparison with the predictive power of hybrid chemical-biological QSAR models (*vide infra*). . The ToxRefDB database includes 78 chronic, carcinogenic, developmental and reproductive toxicity endpoints. However, only 18 out of these were used to develop QSAR toxicity models due to the relatively low fraction of "active"

compounds in the database. The classification Random Forrest (RF), Support Vector Machine with linear kernel (SVM-linear), and Support Vector Machine with RBF kernel (SVM-RBF) QSAR approaches were employed. To ensure the robustness and reliability of the resulting models, we have employed 5-fold cross validation in our modeling process and the same modeling/validation sets were used to develop models with all different machine learning approaches listed above.,. The modeling results show that models for two of these 18 toxicity endpoints yielded Correct Classification Rate (CCR_{ext}) for external compounds consistently above 60% for all three types of QSAR models. After implementing the applicability domain threshold for each individual model, there were four toxicity endpoints, including two chronic and two reproductive rat toxicity endpoints, which have CCR_{ext} higher than 60%. Furthermore, by using specific chemical scaffolds to pre-cluster the original dataset we were able to additionally improve the external predictivity of the resulting model. Thus, our modeling studies suggest that the development of externally validated toxicity predictors, while not without limitations, is feasible for at least some of the ToxRefDB endpoints. We have applied the resultant validated models to screen virtually a chemical library of 50,000 compounds compiled from the European Union's REACH (Registration, Evaluation, Authorization and restriction of CHemical substances) program; we are in the process of consolidating virtual hits, i.e., compounds predicted to be toxic in respective *in vivo* assays. These hist should be regarded as structural hypotheses awaiting their experimental confirmation.

Second, we have employed a novel hierarchical Quantitative Structure-Activity Relationship (QSAR) approach, developed recently in collaboration with Project 2, to develop predictive models for three rat reproductive toxicity endpoints for ToxCast compounds using both chemical and biological (in vitro screening data from ToxCast) descriptors. The prediction accuracy for the best models was in the range of 61-73% for all three *in vivo* endpoints, while that achieved by conventional QSAR models was only 50-65% for the same external set. This suggests that the ToxCast *in vitro* screening data are indeed informative in terms of inferring the prediction of compounds' *in vivo* toxicity. Furthermore, all the ToxCast assays were then ranked based on the external predictivity of the associated models for each *in vivo* toxicity endpoint. Our resulting models could be used to guide the future toxicity studies on the EPA-10K compounds by selecting *in vitro* assays, prioritizing compounds for *in vivo* toxicity evaluation, as well as guiding potential mode-of-action analysis by examining the molecular targets and pathways of the most predictive *in vitro* assays.

To extend our previous successful QSAR modeling efforts using quantitative high-throughput screening (qHTS) data from NTP and NCGC, we aimed at incorporating data on dose-response to chemicals into the modeling routine. Cell viability qHTS data for 1,408 compounds in 13 cell lines have been deposited in PubChem providing the opportunity to study the relationship between *in vitro* and *in vivo effects*. We have identified 400 compounds, for which both qHTS and binary rodent acute toxicity data (i.e., toxic or non-toxic) was also available. We have also developed special noise-eliminating curve fitting procedures to address irregularities in the dose-response curves for some compounds in qHTS data before additional dose-response descriptors are developed for modeling. The external prediction accuracy of conventional QSAR models was 76%. In contrast, the prediction accuracy was above 80% when using hybrid descriptors which account for dose-response. The use of the applicability domain increased prediction accuracy in all models; however, the prediction coverage was decreased only to 81% for the hybrid descriptors, when it fell to 57% for chemical descriptor-based models. These studies suggest that combining qHTS profiles, especially the noise-filtered dose-response qHTS data, with conventional chemical descriptors could potentially improve the predictive power of the computational predictions.

Third, we have employed cheminformatics approaches to the analysis of assertions mined from literature that describe drug-induced liver injury in different species to enable validation of the model predictions. Specifically, in collaboration with Biowisdom

(http://www.biowisdom.com/) we have compiled a dataset of 951 compounds reported to produce a wide range of effects in the liver in different species, comprising humans, rodents, and non-rodents. The liver effects for this dataset were obtained as assertional meta-data, generated from MEDLINE abstracts using a unique combination of lexical and linguistic methods and ontological rules. We have analyzed this dataset using conventional cheminformatics approaches and addressed several questions pertaining to cross-species concordance of liver effects, chemical determinants of liver effects in humans, and the prediction of whether a given compound is likely to cause a liver effect in humans. We found that the concordance of liver effects was relatively low (ca. 39-44%) between different species raising the possibility that species specificity could depend on specific features of chemical structure. Compounds were clustered by their chemical similarity, and similar compounds were examined for the expected similarity of their species-dependent liver effect profiles. In most cases, similar profiles were observed for members of the same cluster, but some compounds appeared as outliers. The outliers were the subject of focused assertion re-generation from MEDLINE, as well as other data sources. In some cases, additional biological assertions were identified which were in line with expectations based on compounds' chemical similarity. The assertions were further converted to binary annotations of underlying chemicals (i.e., liver effect vs. no liver effect), and binary QSAR models were generated to predict whether a compound would be expected to produce liver effects in humans. Despite the apparent heterogeneity of data, models have shown good predictive power assessed by external five-fold cross validation procedures. The external predictive power of binary QSAR models was further confirmed by their application to compounds that were retrieved or studied after the model was developed. To the best of our knowledge, this is the first study for chemical toxicity prediction that applied QSAR modeling and other cheminformatics techniques to observational data generated by the means of automated text mining with limited manual curation, opening up new opportunities for generating and modeling chemical toxicology data.

Finally, in a related effort and in collaboration with project 2, we have applied the QSAR approach to model adverse effects of drugs (AEDs) in liver by constructing binary classification (active vs. inactive) models based on chemical structure. We have employed an FDA's spontaneous reporting database of human liver AEDs (elevations in activity of serum liver enzymes), which contains data on approximately 500 approved drugs. Approximately 200 compounds with wide clinical data coverage, structural similarity and balanced (40/60) active/inactive ratio were selected for modeling and divided into multiple training/test and external validation sets. QSAR models were developed using the k nearest neighbor method and validated using external datasets. Models with high sensitivity (>73%) and specificity (>94%) for prediction of liver AEDs in external validation sets were developed. To test applicability of the models, three chemical databases (World Drug Index, Prestwick Chemical Library, and Biowisdom Liver Intelligence Module) were screened in silico and the validity of predictions was determined, where possible, by comparing model-based classification with assertions in publicly available literature. Validated QSAR models of liver AEDs based on the data from the FDA spontaneous reporting system can be employed as sensitive and specific predictors of AEDs in pre-clinical screening of drug candidates for potential hepatotoxicity in humans.

Activities for Subsequent Reporting Period

In year 3 we plan to continue our emphasis on building predictive toxicity models using the entire chemical structure – in vitro/in vivo data continuum. The ToxCast Phase I results show that there is little direct correlation between *in vitro* assays and *in vivo* endpoints. Thus, there is a need to develop additional methodologies for data organization and data modeling to

improve the predictive ability of associated models. One approach that we will be investigating in collaboration with colleagues in Project 1 is the development of pathway-based biological descriptors. We will use the assay-pathway relationship to map the *in vitro* assay data into different biological sub-spaces represented by individual toxicity pathways before the modeling process. This effort will extend the scope of the traditional chemical-response views to an innovative chemical-pathway-response data landscape and benefit our modeling routine.

In addition, we will begin to focus on using our developing models for virtual screening. The ultimate utility of any computational toxicology method is in its prospective use. Validated QSAR models developed in this project can be applied straightforwardly to evaluate toxicity potential of untested compounds of either environmental or pharmaceutical interest to prioritize candidate molecules that are likely to cause adverse effects for further *in vitro* or *in vivo* studies. This component of the project will include both the annotation of compounds in several publicly available databases, as well as public deposition of our models so that users outside of our laboratory could use them for their compounds of interest. We will continue to make our models publicly available via the ChemBench portal established in our laboratory as a public cheminformatics and toxico-cheminformatics resource (http://chembench.mml.unc.edu/). The users will be able to upload their structures of interest in one of the standard molecular formats (e.g., sdf) and obtain the evaluation of the compounds' toxicity (provided that their compounds will be within the applicability domains of our models). Both annotated compounds in public databases and models should be viewed as major publicly available deliverables of this project.

Publications Arising From this Project

Papers

1. Zhu, H., Ye, L., Richard, A., Golbraikh, A., Wright, F.A., Rusyn, I., and Tropsha, A. (2009) A novel two-step hierarchical quantitative structure activity relationship modeling workflow for predicting acute toxicity of chemicals in rodents. Envr Health Persp 117:1257-1264 (*from in press to published*).
2. Zhu, H., Martin, T.M., Ye, L., Young, D.M., and Tropsha, A. (2009) Combinatorial QSAR modeling of rat acute toxicity by oral exposure. Chem Res Toxicol 22:1913-1921 *(from submitted to published).*
3. Fourches, D., Barnes, J.C., Day, N.C., Bradley, P., Reed, J.Z., and Tropsha, A. (2010) Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. Chem Res Toxicol 23:171-183.
4. Rodgers, A.D., Zhu, H., Fourches, D., Rusyn, I., and Tropsha, A. (2010) Modeling Liver-Related Adverse Effects of Drugs Using kNN QSAR Method. Chem Res Toxicol 23:724-732 (*from submitted to published*).

Posters/Abstracts

1. Zhu, H., Tropsha A. (2009) The use of hybrid chemical/biological descriptors in QSAR modeling improves the accuracy of in vivo chemical toxicity prediction, The 32nd Annual Midwest Biopharmaceutical Statistics Workshop, Muncie, IN.
2. Zhu, H., Sedykh, A., Wright, F., Rusyn, I., Tropsha, A. (2009) Using quantitative High Throughput Screening (q-HTS) results as biological descriptors to assist modeling of acute rat toxicity, American Chemical Society 238th National Meeting, Washington, DC.
3. Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., and Tropsha, A. (2010) Using *in vitro* Dose-Response Profiles to Enhance QSAR Modeling of *in vivo* Toxicity. Society of Toxicology Annual Meeting, Salt Lake City, UT.