**Preliminary Data and Work Progress**

This year, the main efforts of Project 3 was devoted to further development of Quantitative Structure-Activity Relationship (QSAR) methodology and its applications to chemical toxicity datasets. The gist of our approach (that may be gleaned even from the title of our project) is the joint use of both chemical descriptors of molecules and the results of their biological testing *in vitro* as hybrid *chemical biological descriptors* in our predictive QSAR modeling workflow. The details of our recent methodological advances are outlined in the Results up to date section below and in recent publications cited therein. Armed with the aforementioned methodologies, we have recently initiated the analysis of ToxCast data that were released early February. We have submitted two poster abstracts for the upcoming EPA Workshop in May 2009 and will be making an oral presentation as well. In one of these studies we have employed cheminformatics approaches to analyze the EPA ToxCast chemical libraries to develop predictive toxicity models and prioritize compounds for *in vivo* toxicity testing. In another study we focused on a fraction of ToxCast assays that reported the largest numbers of compounds "active" in both *in vitro* and *in vivo* assays. We have employed ToxCast cell-viability and gene-expression assays as biological descriptors in QSAR modeling of animal toxicity endpoints. Specifically, we have analyzed the relationship between the short-term bioassay results and different chronic *in vivo* toxicity responses for the ToxCast Phase I compounds.

**Results to Date**

A wealth of available biological data requires new computational approaches to link chemical structure, in vitro data, and potential adverse health effects. We have continued to advance the predictive QSAR modeling workflow that relies on effective statistical modle validation routines and implements both chemical and biological (i.e., *in vitro* assay results) descriptors of molecules to develop *in vivo* chemical toxicity models. In collaboration with project 2, our recent study that was just accepted as a paper by EHP (Zhu et al, 2009, in press) has explored a database containing experimental *in vitro* $IC_{50}$ cytotoxicity values and *in vivo* rodent $LD_{50}$ values for more than 300 chemicals that was compiled by the German Center for the Documentation and Validation of Alternative Methods (ZEBET). The application of conventional Quantitative Structure-Activity Relationship (QSAR) modeling approaches to predict mouse or rat acute $LD_{50}$ from chemical descriptors of ZEBET compounds yielded no statistically significant models. The analysis of these data showed no general significant correlation between $IC_{50}$ and $LD_{50}$. However, a linear $IC_{50}$ vs. $LD_{50}$ correlation could be established for a fraction of compounds. To capitalize on this observation, a novel two-step modeling approach was developed as follows. First, all chemicals are partitioned into two groups based on the relationship between $IC_{50}$ and $LD_{50}$ values: one group is formed by compounds with linear $IC_{50}$ vs. $LD_{50}$ relationship, and another group consists of the remaining compounds. Second, conventional binary classification QSAR models are built to predict the group affiliation based on chemical descriptors only. Third, k-Nearest Neighbor continuous QSAR models are developed for each sub-class to predict $LD_{50}$ from chemical descriptors. All models have been extensively validated using special protocols. The novelty of this modeling approach is that it uses the relationships between in vivo and in vitro data only to inform the initial construction of the hierarchical two-step QSAR models. Models resulting from this approach employ chemical descriptors only for external prediction of acute rodent toxicity. A significant result of this study is that not only did we show that the use of this hybrid modeling approach yields more accurate models of in vivo rodent chemical toxicity but we have also demonstrated that our models are superior to commercial approaches such as TOPKAT when applied to the same group of chemicals.

In another recent study in collaboration with scientists from EPA-Cincinnati we have focused on rodent toxicity. Few QSAR studies have successfully modeled large, diverse rodent toxicity endpoints. In our study, a combinatorial QSAR approach has been developed for the creation of robust and predictive models of acute toxicity in rats caused by oral exposure to chemicals. To this end, a comprehensive dataset of 7,385 compounds with their most conservative lethal dose ($LD_{50}$) values has been compiled. To evaluate the predictive power of the models generated in this research, a comparison was undertaken with an available commercial toxicity predictor, TOPKAT. A modeling set was developed that included all of the 3,472 compounds that TOPKAT used in its modeling set. The remaining 3,913 compounds which were not present in the TOPKAT modeling set were used as the external validation set for our toxicity models. Five different types of QSAR $LD_{50}$ models were developed for the modeling set. The prediction accuracy for the external validation set ranged from 0.24 to 0.70 (linear regression coefficient $R^2$). The use of applicability domain threshold implemented in most models generally improved the external prediction accuracy but obviously led to the decrease in chemical space coverage. Ultimately, several consensus models were developed by averaging the predicted $LD_{50}$ for every compound using all 5 models. The consensus models afforded higher prediction accuracy for the external validation dataset with the highest coverage as compared to individual constituent models. The validated consensus $LD_{50}$ models developed in this research can be used as reliable computational predictors of *in vivo* acute toxicity and will be made publicly available from the participating laboratories. The paper (Zhu et al, 2009) has been submitted to Chemical Research in Toxicology.

In addition, the Hierarchical Technology for QSAR was applied to 95 various nitroaromatic compounds (including some widely known explosives) tested for their toxicity (50% inhibition growth concentration, $IGC_{50}$) against the ciliate *Tetrahymena pyriformis.* The dataset was divided into multiple training and test sets according to the putative mechanisms of toxicity and multiple models have been generated with leave-one-out cross-validated $Q^2$=0.68–0.86 and $R^2$=0.82–0.95 for the training sets and $R^2$ 0.57-0.86 for the test sets. The resulting models were employed for predicting toxicity of an external set of 63 nitroaromatics and $R^2_{ext}$=0.65 was obtained. Validated models were explored to (i) elucidate the effects of different substituents in nitroaromatic compounds on toxicity; (ii) differentiate compounds by probable mechanisms of toxicity based on their structural descriptors ; (iii) analyze the role of various physical-chemical factors (e.g., electrostatics, hydrophobicity, hydrogen bonding, atomic identity, etc.) in compounds' toxicity. Molecular fragments favoring or interfering with toxicity were defined by interpretation of obtained PLS models. For example, oxibutane and aminophenyl substituents promote toxicity of nitroaromatics against *Tetrahymena Pyriformis* but carboxyl group decreases toxicity. It was also shown that mutual influence of substituents in benzene ring plays the determining role in toxicity variation. In addition, the Classification and Regression Trees (CART) method has been used successfully for generating models that predict possible mechanism of compound toxicity. The prediction accuracies of CART models for test set compounds ranged between 79 and 84%.

We are also actively working with the ToxCast Phase I data. First, we have employed conventional QSAR modeling approaches utilizing chemical descriptors of pesticides only to in an attempt to develop predictive models of *in vivo* toxicity; however, no statistically significant and externally predictive models for the EPA-320 set resulted from these efforts. We then employed a variable QSAR approach that combines the results of selected ToxCast *in vitro* bioassays as additional biological descriptors and a subset of chemical descriptors into hybrid chemical and biological descriptors of pesticides. We have found that the predictivity of these models for *in vivo* toxicity prediction has improved significantly. Thus, our modeling studies suggest that a subset of ToxCast assays in combination with selected chemical descriptors are indeed informative of compounds' in vivo toxicity. We expect that our models could guide future experimental in vitro studies of the EPA-10K compounds towards their focused *in vitro* testing and prioritization for *in vivo* toxicity evaluation.

Second, we focus on a fraction of ToxCast assays that reported the largest numbers of compounds "active" in both *in vitro* and *in vivo* assays. We have employed ToxCast cell-viability and gene-expression assays as biological descriptors in QSAR modeling of animal toxicity endpoints. Specifically, we have analyzed the relationship between the short-term bioassay results and different chronic *in vivo* toxicity responses for the ToxCast Phase I compounds. First, we have transformed the data such as to reduce the noise associated with the ToxCast bioassay results and make the data more suitable for QSAR modeling. Second, by using statistical tools developed in house, we have selected a subset of bioassays that we regarded as potentially important for explaining *in vivo* toxicity of ToxCast compounds. Third, all the ToxCast compounds were assigned to several classes based on their biological response patterns in the selected *in vitro* and *in vivo* assays. This reformatted *in vitro - in vivo* toxicity dataset was used to develop classification QSAR models using the *k* Nearest Neighbor (*k*NN) method and Dragon descriptors. The resulting QSAR models could be used to assign new compounds to the appropriate *in vitro- in vivo* correlation class and then predict the associated *in vivo* toxicity. Utilizing *in vitro* assay results as biological profiles makes this modeling workflow superior to the conventional QSAR modeling that utilizes chemical descriptors only.

**Activities for Subsequent Reporting Period**

Activities for next year in Project 3 will focus primarily on the analysis of ToxCast data, but will also continue to explore other datasets that provide both *in vivo* and *in vitro* data for chemicals. For ToxCast data analysis our primary goal is to build models that could be used by EPA to prioritize the selection of Phase 2 compounds. In this endeavor we rely on our predictive QSAR modeling workflow that incorporated a module for virtual screening of external chemical libraries. Thus, after completing the first cycle of model building for ToxCast Phase I data we will employ our models to virtually screen the 10,000 compounds in the EPA-10K list. So far, we have developed two distinct methodologies for *in vivo* toxicity prediction utilizing both chemical and biological descriptors. In the first approach, we employ biological descriptors directly in combination with chemical descriptors to build models. Obviously, this approach requires the knowledge of biological descriptors to make toxicity assessment for new compounds. We will explore approaches that could help circumvent this relative limitation of our method by predicting *in vitro* toxicity profiles from conventional chemical descriptors based QSAR models that are built independently using *in vitro* data as the target property.

Our second modeling approach employs the explicit relationship between *in vitro* and *in vivo* data as part of the two-step hierarchical modeling strategy where a special binary QSAR model using chemical descriptors only is built to partition compounds into classes defined by patterns of *in vitro – in viv*o relationships. In the second step, class specific conventional QSAR models are built, also using chemical descriptors only. Thus, this hierarchical strategy affords external predictions using chemical descriptors only. Again, the ToxCast models built with this approach will be utilized for prioritizing the EPA-10K compounds for Phase 2 testing.

In addition to the above plans, we will also continue making progress in both methodology development and applied studies using additional datasets. We are currently working on improved definition of model applicability domains, which is critical to enable accurate prediction of the largest possible external dataset. We continue to integrate novel data analytical approaches into our QSAR modeling workflow. For instance, we have been exploring a data mining methodology termed Complete Correlation Analysis that appears most useful when comparing two data spaces that are expected to be related to each other such as arrays of *in vivo* and *in vitro* assays.  We also intend to complete the second round of studies into structure – *in vitro* cytotoxicity – *in vivo* carcinogenicity relationships using extended dataset generated in collaboration between EPA, National Toxicology Program, and National Chemical Genomic Center at NIH.

## Publications Arising From this Project

*Papers*

1. Zhu H, Rusyn I, Richard A, Tropsha A. (2008) Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. Environ. Health Perspect. 116: 506-513.
2. Zhu H, Ye L, Richard A, Golbraikh A, Wright FA, 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 (*In Press*)
3. Zhu H, Martin TM, Ye L, Young DM, and Tropsha A. (2009) Combinatorial QSAR modeling of rat acute toxicity by oral exposure. Chem Res. Tox. (*Submitted*)
4. Artemenko AG, Muratov EN, Kuz'min VE, Gorb LG, Quasim M, Leszczynski J, and Tropsha A. (2009) QSAR analysis of structural factors and possible modes of nitroaromatics' toxicity in *Tetrahymena Pyriformis.* Chem Res. Tox. *(Submitted)*

*Posters/Abstracts*

1. Zhang L, Zhu H, Rusyn I, Judson R, Dix D, Houck K, Martin M, Richard A, Kavlock R, and Tropsha A. (2009) Cheminformatics Analysis of EPA ToxCast chemical libraries to identify domains of applicability for predictive toxicity models and prioritize compounds for toxicity testing. Society of Toxicology Annual Meeting, Baltimore, MD.
2. Zhu H, Martin TM, Ye L, Young DM, and Tropsha A. (2009) Combinatorial QSAR Modeling of Rat Acute Toxicity by Oral Exposure. Society of Toxicology Annual Meeting, Baltimore, MD.
3. Easterling M, Tropsha A, Ye L, Pirone J, Zhu H, Martin T, and Moudgal C. (2009) Quantitative Structure Toxicity Relationships (QSTR) models for predicting acute/sub-acute and sub-chronic/chronic adverse effect levels. Society of Toxicology Annual Meeting, Baltimore, MD.
4. Zhang L, Judson R, Dix D, Houck K, Martin M, Richard A, Kavlock R, and Tropsha A. (2009). Cheminformatics Analysis of EPA ToxCast chemical libraries to develop predictive toxicity models and prioritize compounds for *in vivo* toxicity testing. EPA ToxCast Data Analysis Summit, Durham, USA.
5. Zhu H, Sedykh A, Zhang L, Rusyn I, and Tropsha A. (2009). Using ToxCast cell-viability and gene-expression assays as biological descriptors in QSAR modeling of animal toxicity endpoints. EPA ToxCast Data Analysis Summit, Durham, USA.
6. Wang K, Golbraikh A, and Tropsha A. (2009). Optimization of pattern recognition and classification by combinatorial QSAR modeling of the carcinogenic potency database. 237th ACS National Meeting, Salt Lake City.
7. Kuz'min V, Muratov E, Varlamova E, Artemenko A, Kovdienko N, and Tropsha A. (2009) QSTR analysis of mixtures toxicity to Daphnia magna. 237th ACS National Meeting, Salt Lake City.