Carolina Center for Computational Toxicology

Funded by U.S. EPA Cooperative Agreement STAR RD 833825
Progress Report for Project Period 04/01/11 - 03/31/12
Report Date: 02/19/12
Ivan Rusyn, Principal Investigator

The Carolina Center for Computational Toxicology (CCCT) is comprised of three research projects and an administrative core. The major aims and objectives of the CCCT have not changed from the original application. The content of this progress report is organized according to U.S. EPA guidelines and it summarizes significant activities and accomplishments of all four components of the CCCT.

Preliminary Data and Work Progress

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

In the current year of the project we have continued to develop new models of various toxicity endpoints using, when available, short term assay data under the general paradigm of chemical structure – *in vitro* – *in vivo* extrapolation. Under the rubric of "conventional" QSAR modeling, novel QSAR models of skin sensitization using chemical descriptors of compounds only have been developed and validated; the predictive power of these models was shown to exceed significantly that of predictions made with read-across method as implemented within the OECD QSAR Toolbox. In another study, we have developed new QSAR models of endocrine disrupting chemicals (EDCs), with a unique caveat that the large datasets of compounds interacting with both ERα and/or ERβ were collected from public databases and scientific literature. At the same time, we have made significant progress in developing new approaches to modeling *in vivo* toxicity using both chemical descriptors and short term assays results. In the previous years of the grant, we have established that short-term assay results treated as special biological descriptors of the same chemicals when used in combination with chemical descriptors afford hybrid models exceeding traditional QSAR models in terms of their external predictive power. We have summarized our hybrid approaches and classified them in three different categories. We have also extended our previous studies looking into the use of microarray gene expression data as biological descriptors that can be combined with conventional chemical descriptors to predict hepatotoxicity. Our previous published studies demonstrated that toxicogenomics descriptors afforded models with higher predictive power than those using either chemical descriptors alone or chemical descriptors in combination with toxicogenomics descriptors. However, in the current year of the grant we have developed a new approach termed hybrid read-across where compound hepatotoxicity (i.e., toxic or non-toxic) is predicted to match that of the majority of molecules with both similar chemical structure, as well as with similar toxicogenomic expression profile. Additional studies are in progress testing this new method on additional datasets (such as the Iconix/CEBC dataset) as well as using more sophisticated approaches for identifying chemically and/or biologically similar compounds to improve upon the current implementation of the hybrid read-across method.

Results to Date

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

Since QSAR approaches continue to be central to our research, we have continued to apply best practices in the field to new datasets of compounds tested against important toxicity endpoints that were collected from the literature or electronic databases. We have developed QSAR models of skin sensitization (tested in local lymph node assay) and compared their performance in terms of external predicitivity with the models developed using read across method (as implemented in the OECD QSAR toolbox). We have applied both kNN algorithm, which has been widely investigated in our group, and RF classification algorithm to a dataset of 471 compounds, which was obtained from the 2009 annual report of Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) program in NIEHS. The results showed that the sensitivity, specificity and correct classification rate (CCR) for external validation dataset were 89%, 69% and 79% for kNN models and 81%, 73% and 77% for RF models. Both kNN and RF models have explicitly incorporated the applicability domain. Furthermore, result of y-randomization and 5-fold external validation demonstrated the robustness and stability of the QSAR models. We also applied the OECD QSAR toolbox to evaluate skin sensitization potential of a group of compounds used for external validation of our QSAR models that also were not included in the training set within the OECD toolbox, and the accuracy of prediction was only ca. 50%. The results showed significant advantage of our QSAR models over the OECD toolbox in terms of predictive accuracy. Thus, our QSAR models could be used as reliable predictors of the skin sensitization potential of chemicals.

In another QSAR modeling study, we have analyzed endocrine disrupting chemicals (EDCs) that have become a growing public concern due to their adverse effects on human and wild life. In order to develop *in silico* predictors to identify Estrogen Receptor (ER)-mediated EDCs, a large number of ER ligands were collected from public databases and scientific literature, with relative binding affinity to ERα and/or ERβ (546 compounds for ERα and 137 compounds for ERβ). A novel multi-task learning (MTL) QSAR modeling approach was applied to develop models capable of predicting the binding affinity of ligands to both ER subtypes. Compared with conventional single-task learning (STL) models, MTL models significantly improved the predictive accuracy for ERβ binding affinity (R^2 increased from 0.32 to 0.53) while keeping the high predictive accuracy of ERα models (MTL R^2=0.71 vs. STL R^2=0.73). In addition, as a complementary approach, docking studies were performed on a set of ER agonists/antagonists (67 agonists/39 antagonists for ERα and 48/32 for ERβ) and corresponding presumed decoys/non-binders (2570/1448 for ERα and 1000/1000 for ERβ). These compounds were docked to four protein conformations: ERα agonist (PDB ID: 1L2I), ERα antagonist (PDB ID: 3DT3), ERβ agonist (PDB ID: 2NV7), and ERβ antagonist (PDB ID: 1L2J), respectively. Results showed that all four conformations were capable of discriminating their corresponding ligands from presumed decoys/non-binders, with ERα agonist conformation being the best in separating ERα agonists from antagonists. VS of an uterotrophic dataset validated that the consensus of MTL QSAR and docking models had the highest enrichment power. Virtual screening of the EPA Tox21 library yielded a prioritized list of 286 putative estrogenic compounds for future *in vitro* and *in vivo* tests on endocrine disruption.

In continuation of our research focused on the development of hybrid approaches using both chemical and biological (short term assays) compound descriptors, we have continued our analysis and modeling of a compound library screened at the NIH Chemical Genomics Center (NCGC). These studies have been done in collaboration with our colleagues working on Project 2. As discussed in the Project 2 report population-based toxicity screening has been conducted against 81 cell lines to test a dataset comprised of 240 compounds selected from the 1,408

substances of the NTP screening library. Each compound was tested in each cell line using cell viability and caspase assays. The quality control analysis reported last year demonstrated high reproducibility for both assays. Consequently, we have applied these qHTS data in our standard modeling workflow to develop rigorous models of three *in vivo* endpoints: rat acute toxicity, Ames mutagenicity, and rodent carcinogenicity. In tune with the previous studies we have developed models using both conventional chemical descriptors as well as noise-treated qHTS data employed as biological descriptors. Biological descriptors alone afforded acceptable models, which were as good as conventional QSAR models (accuracy 54-65% for both types), while hybrid models, employing biological and chemical descriptors, achieved accuracy of up to ~70%. For comparison, similar modeling studies employing cytotoxicity qHTS data from 13 NCGC cell-lines yielded inferior biological (accuracy 46-58%) and hybrid (accuracy 56-65%) models. Therefore, we conclude that a screening panel of 81 lymphoblast cell lines is likely to have higher statistical power for predicting in vivo effects. Interestingly, among the three modeled *in vivo* endpoints, carcinogenicity was most accurately predicted by biological models with 65% accuracy and positive predictive value of 72%. This indicates that pure biological models based on short-term cytotoxicity data have a potential to become effective prioritizing tools for such long-term in vivo endpoints as rodent carcinogenicity. Furthermore, we have analyzed the 240 compounds in terms their structural diversity as well as diversity of their *in vitro* biological response in order to understand the possible reasons for the observed trends in the prediction accuracy of the models. We are also investigating the relationships between chemical structure and cell-specificity of biological response with the goal of interpreting obtained models and obtaining recommendations for the most informative cell lines. At the next stage, we plan to include into our analysis and into the hybrid modeling workflow the genotype information available in form of SNPs data for 74 of the screened lymphoblast cell lines.

Finally, we have continued to explore the power of chemical descriptors and toxicogenomics profiles for predicting hepatotoxicity. Last year, we reported on our analysis of the Japanese Toxicogenomics Project where rat liver microarray database of 127 drugs was examined (http://toxico.nibio.go.jp/datalist.html). The target property was rat hepatotoxicity following 28 days of treatment, established by histopathology and serum chemistry. Last year, we reported that toxicogenomics models using only transcriptional profiles (76% Correct Classification Rate, CCR) outperformed conventional QSAR models (61% CCR, descriptors used: Dragon; Molecular Operating Environment, MOE; and simplex representation of molecular structure, SiRMS) and hybrid models using both chemical and toxicogenomic descriptors (60-77% CCR). During the current year of the grant, we have developed a novel multi-space *k* nearest neighbors read-across (MSKRA) method that resulted in the best prediction performance to date (79% CCR). To describe the method briefly, it is based on the popular and transparent read-across approach implemented using cheminformatics concept of global similarity in the space of multiple chemical descriptors. Under this implementation, *k* nearest neighbors of a test compound are independently identified in both the chemical and toxicogenomic descriptor spaces using Tanimoto similarity (forming a group of 2k neighbors) and the predicted toxicity is computed from the similarity-weighted average of toxicities of all 2*k* neighbors. MSKRA also afforded the best model for another similar data set, Iconix, containing both toxicogenomics and chemical descriptors. Our studies demonstrate that with further method development, there is a strong advantage in combining chemical descriptors with biological assays to improve the predictivity and interpretability of computational toxicology models.

Activities for Subsequent Reporting Period

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

In the next year, we plan to place major focus on enhancing our hybrid approaches to *in vivo* toxicity prediction using both chemical (computed from chemical structure) and biological (generated in short term in vitro or in vivo assays) descriptors. The computational approach will continue to be QSAR modeling, and we do expect to develop conventional QSAR models of selected toxicity endpoints (e.g., Ames genotoxicity). In addition, we will be developing conventional QSAR models of any end point (especially those in ToxRefDB) for which short term assays results are available for building hybrid models. As far as novel theoretical developments, we will continue to explore multi-task learning methods especially for predicting results of related in vivo toxicity assays, e.g., those characterizing biological response to chemicals at the organ or tissue levels (e.g., all assays relevant to hepatotoxicity) or chemical toxicity effects mediated by the same pathway. With respect to specific applications, we intend to spend significant effort on the analysis of ToxCast Phase II. This new data collection that will provide the in vitro assay results for nearly 1000 chemicals will be ideal for our extensive application of hybrid modeling technologies to advance the application of our general chemical structure- in vitro – in vivo extrapolation approach. In addition, we will be working with Projects 1 and 2 to provide cheminformatics component towards the joint analysis of the data from the current experiment on 1100+ cell lines and ~200 chemicals performed in partnership with Tox21.

Publications Arising From this Project in Year 4

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

Papers

1. Golbraikh, A., Wang, X.S., Zhu, H., and Tropsha, A. (2012) Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment. In: Handbook of Computational Chemistry, Shukla and Leshinkii, Eds, Springer (in press).
2. Low, Y., Uehara, T., Minowa, Y., Yamada, H., Ohno, Y., Urushidani, T., Sedykh, A., Muratov, E., Kuz'min, V., Fourches, D., Zhu, H., Rusyn, I., Tropsha, A. (2011) Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. Chem Res Toxicol. 24(8):1251-62. (*from submitted to published*).
3. Rusyn, I., Sedykh, A., Low, Y., Guyton, K.Z., Tropsha, A. (2012) Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. Toxicol. Sci. (*in press*)

Posters/Abstracts

1. Wignall, J., Sedykh, A., Tropsha, A., Woodruff, T. J., Zeise, L., Rusyn, I., Cogliano, V. J., Chiu, W.A., Guyton, K. Z. Modeling Toxicity Values Using Chemical Structure, In Vitro Screening, And In Vivo Toxicity Data. Society of Toxicology Annual Meeting, San Francisco, CA, 2012.
2. Low, Y., Fourches, D., Sedykh, A., Rusyn, I., Tropsha, A. Multi-space k-Nearest Neighbors as a Novel Hybrid Approach Integrating Chemical and Toxicogenomic

Descriptors for Improved Toxicity Prediction. Society of Toxicology Annual Meeting, San Francisco, CA, 2012.
3. Sedykh, A., Low, Y., Lock, E., Rusyn, I., Tropsha, A. Using population-based dose-response cytotoxicity data for *in silico* prediction of rat acute toxicity. Society of Toxicology Annual Meeting, San Francisco, CA, 2012.