Modern In Vitro Imputation Modeling and Machine Learning to Predict Chemical Carcinogenicity

A. Borrel[1], G. Tedla[1], K. T. To[1], B. Hill[1], A. Karmaus[2], A. Wang[3], T. Luechtefeld[4], D. Reif[5], D.G. Allen[1], N. Kleinstreuer[6]

[1]Inotiv, RTP, NC, United States; [2]Syngenta, Greensboro, NC, United States; [3]NIH/NIEHS/DTT/IHAB, RTP, NC, United States; [4]Insilica LLC, Bethesda, MD, United States; [5]NIH/NIEHS/DTT/PTB, RTP, NC, United States; [6]NIH/NIEHS/DTT/NICEATM, RTP, NC, United States

Background and Purpose

Carcinogenesis is a multistep process in which normal cells acquire properties that allow them to form either malignant or non-malignant cancers. The concept of the ten key characteristics of carcinogens (KCC) was developed by evaluating attributes of known chemicals and viruses that induce human cancers. Quantitative structure-activity relationship (QSAR) models that rely on structural or physicochemical properties to predict carcinogenesis potential as an apical endpoint lack sufficient information on the complex multiple mechanisms involved in carcinogenicity. This project addresses that lack of readily computable mechanistic information via combining chemical-specific prediction of KCC potential with a novel machine learning-based imputation approach to account for missing data from Tox21/ToxCast in vitro high-throughput screening (HTS) assays.

Methods

We imputed missing activities for a data set of 10,000 chemicals and 2,000 in vitro HTS assay endpoints. Initially, we built imputation models using structural and physicochemical properties in addition to the available bioactivity information. Subsequently, we enriched the in vitro data from Tox21/ToxCast by leveraging the resources from BioBricks.ai, a bioinformatics inventory that compiles toxicity-relevant databases into a harmonized and easily accessible format. This enrichment enabled us to include additional information such as protein target binding into the imputation model. We tested up seven different machine learning approaches for imputation and assessed performance on a random subset of 15% of the measured activities. Finally, using the imputed data, we leveraged ToxPi (https://toxpi.org) to score each chemical's carcinogenicity likelihood by KCC. Carcinogenicity expert input was applied to map the ToxCast/Tox21 assays onto the KCC. This mapping is available within NICEATM's Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/).

Results

We found that the ExtraRegressor machine learning approach performed best for the imputation, with test set $R^2 > 0.8$. We discovered that the imputation modeling outperformed any classic QSAR model by at least a 0.2 increase in the R-squared value on an exemplar set of selected assays. Increasing the amount of data in the imputation modeling by using repositories in BioBricks.ai did not change performance of the imputation but allowed us to improve KCC scoring. Using imputed data, we constructed a profile for each chemical which included a score between 0 and 1 for each KCC. We validated this approach by examining carcinogens that were identified as active on one or several KCCs by authoritative agencies such as the International

Agency for Research on Cancer Monograph Program. We found that we were able to accurately predict the mapped KCC for well-studied carcinogens such as lindane and bis(chloromethyl) ether.

Conclusions

In this work, we developed valuable resources for carcinogenicity research, including a consensus list of carcinogenic chemicals from various U.S. and international agencies as well as an updated mapping of KCC onto ToxCast/Tox21 assays. We leveraged modern artificial intelligence imputation modeling and machine learning models to predict chemical carcinogenicity profiles via the KCC, which we subsequently validated using reference carcinogens. This work opens numerous possibilities for utilizing imputation modeling to address various toxicologically relevant endpoints for which sparse data from multiple sources are available.

This project received funding from federal funds provided by the NIEHS, NIH under Contract No. HHSN273201500010C.

Keywords: AI, carcinogenicity, Tox21, In vitro assays, QSAR