Domain-Specific QSAR Model for Identifying Potential Estrogenic Activity

K Mansouri[1], R Politi[2], S Farag[2], E Muratov[2], A Tropsha[2], R Judson[1], Q Zang[3], D Allen[3], W Casey[4], N Kleinstreuer[3]

[1]EPA/ORD/NCCT, RTP, NC, USA, [2]UNC-CH, Chapel Hill, NC, USA; [3]ILS/NICEATM, RTP, NC, USA; [4]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

Humans are potentially exposed to tens of thousands of man-made chemicals in the environment, some of which may mimic natural endocrine hormones and thus have the potential to be endocrine disruptors. Predictive *in silico* tools can be used to quickly and efficiently evaluate these untested chemicals for their ability to disrupt the endocrine system. The Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) brought together international collaborations to build an ensemble of models to identify chemicals with the potential to interact with the estrogen receptor (ER). Various QSAR models from different groups were trained and validated on the ToxCast/Tox21 ER assay data, and then used individually and in combination to screen a large library of ~30,000 chemicals for ER binding and agonist activity. The CERAPP results showed a high prevalence of phenolic compounds in the set of predicted positives, consistent with prior knowledge on the influence of this structural moiety on chemical interaction with the ER. However, because CERAPP global models did not accurately predict specific activity and relative potency of various phenols, we constructed local QSAR models focused on this chemical category. Phenolic compounds were identified in the curated CERAPP dataset by the presence of a benzene ring with a hydroxyl substituent (aryl ethers were not considered.) Various machine learning approaches, such as random forest and partial least squares discriminant analysis, were applied to build local QSAR models for ER binding and agonist activity of phenolic compounds. These models were trained and tested on data from the ToxCast/Tox21 assays and well-curated literature sources with independently reproducible results. The local models consistently yielded higher predictivity (balanced accuracies ~0.9) and better balance between sensitivity and specificity than the global models as evaluated by their performance on the external test sets. These models can be used as regulatory decision support tools for evaluating the endocrine disrupting potential of environmental chemicals. *This work does not reflect EPA policy. This project was funded in whole or in part with Federal funds from the NIEHS, NIH under Contract No. HHSN273201500010C*.