A Systematic Evaluation of Analogs and Automated Read-across Prediction of Estrogenicity

Prachi Pradeep[1,2], Kamel Mansouri[1,2], Grace Patlewicz[2] and Richard Judson[2]

[1]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee
[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Science and Decision Context

Read-across is a data gap filling technique widely used within category and analog approaches whereby property information for one or more chemicals (source analogs) is used to predict the same property for a data-poor (target) chemical, which is considered to be "similar" in some way. Source analogs are typically identified on the basis of structural similarity. Although much technical guidance has been published for read-across, practical principles for identification and evaluation of the scientific validity of source analogs is still lacking. This case study sought to investigate (1) the ability of three structure descriptor methods (Pubchem, Chemotyper and MoSS) to identify analogs for read-across and predict Estrogen Receptor (ER) binding activity, and (2) the utility of data quality measures, physicochemical properties, and R-group properties for filtering relevant analogs to ascertain better predictions and improvement in uncertainty associated with read-across ER binding predictions, for a specific class of chemicals: hindered phenols. Hindered phenols are phenols with one or more bulky functional groups ortho to the hydroxyl group. E.g. 3-Chloro-4-hydroxybenzoic acid: ()

The dataset comprised 462 hindered phenols and 257 non-hindered phenols. The results demonstrate that: (1) concordance in ER activity increases with similarity, (2) data quality significantly improves read-across predictions, and (3) filtering analogs using global and local properties results in more relevant analogs for read-across predictions.

This case study illustrates that the quality of experimental data and use of biologically-relevant chemical descriptors to identify source analogs are critical to a robust read-across prediction.

Approach

1. Structural source analogs were identified using 3 different chemical structure descriptor approaches (Pubchem, Chemotyper and MoSS MCSS) and Tanimoto index as a measure of similarity
2. Concordance analysis and a read-across ER binding prediction was done for each target hindered phenol

ANALOG SELECTION METHOD

Descriptor Approach	Basis
Pubchem	881 bits fingerprints
MoSS MCSS	Size of most common substructure
ToxPrints/Chemotyper	Chemical substructures fingerprint with pre-defined Chemotypes

Underlying basis for each of the three chemical descriptor approaches

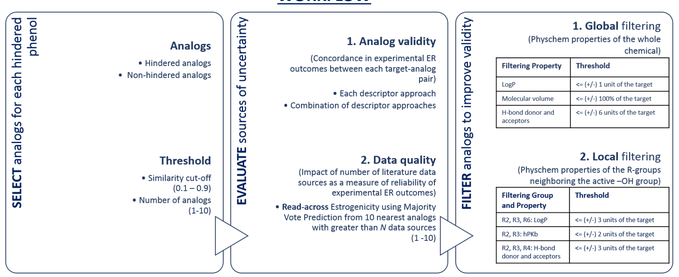
DATASET

Curated data set from different over lapping sources including: Tox21, FDAEDKB, METI database, ChEMBL and other sources from CERAPP project.

Target: 462 hindered phenols
Inventory of Source Analogs: 719

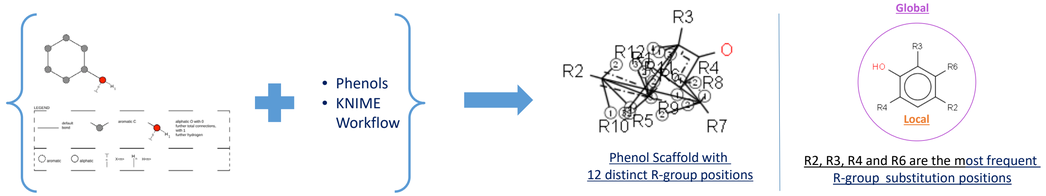
Target phenols (≥ 4 data sources): 296
Source Analogs (≥ 4 data sources): 481

WORKFLOW



Results

R-GROUP DECOMPOSITION



The phenol scaffold and KNIME workflow was used to decompose each phenol into R-group substituents

UNCERTAINTY ANALYSIS

1. Data Quality
(N: No. of analogs, T: No. of hindered phenols predicted)

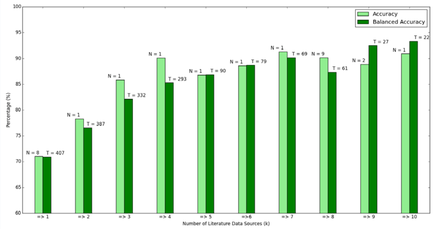
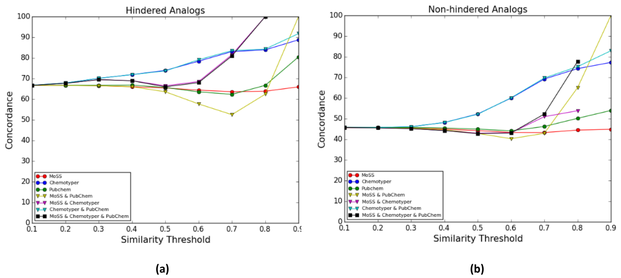


Figure 1: Literature data source analysis to observe the effect of data quality on read-across predictions. The x-axis corresponds to the threshold in number of data sources and the y-axis corresponds to the maximum accuracy/balanced accuracy of prediction for the dataset. The text on top of each bar plot indicates the number of analogs resulting in the best prediction (N) and the number of hindered phenols that had at least N analogs (i.e. were predicted) from the restricted dataset (T).

2. Concordance
(Data: Phenols with ≥ 4 data sources)



(a) (b)

Figure 2: Concordance analysis using phenols with ≥ 4 literature data sources. The dataset comprises 298 hindered phenols and 183 non-hindered phenols. (a) Using hindered phenols as analogs , and (b) Using non-hindered phenols as analogs for each target hindered phenol.

READ-ACROSS RESULTS

Method: PubChem
Data: Phenols with ≥ 4 data sources

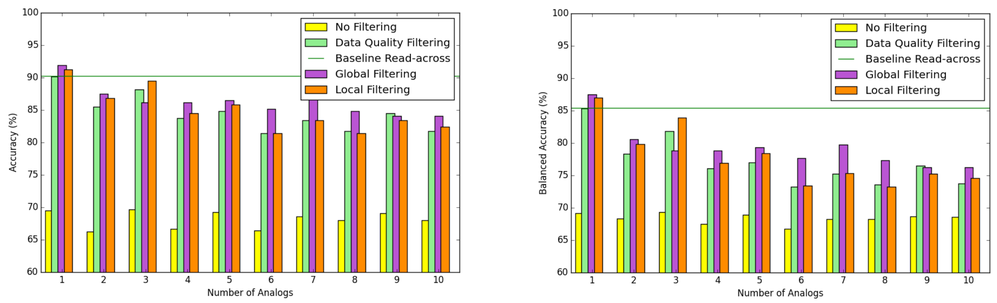


Figure 3: Accuracy and balanced accuracy: effect of global and local filtering on analog quality and read-across predictions.

Anticipated Impacts

Read-across is a conceptually simple and scientifically sound technique. However, identification of relevant and valid analogs for read-across prediction for any endpoint is not trivial. This case study illustrates that the quality of experimental data and use of biologically-relevant chemical descriptors to identify source analogs are critical to a robust read-across prediction.

1. Concordance analysis for each descriptor method using each target-analog pair (with a similarity cut-off) indicates that the concordance in ER activity rises with increasing similarity
2. (a). Data quality analysis illustrates the importance of using good data (validated from multiple sources) and its impact in reducing uncertainty in the quality of read-across predictions. Setting limits on data source thresholds drastically improves prediction accuracy
(b). Filtering of analogs based on conceptually simple steric and electronic properties improves the validity of analogs and subsequently prediction accuracy

Using only one (nearest) analog with good quality data, performs as well as any other combination (balanced or total accuracy). This provides support for using the standard "analog" approach in read-across.

ILLUSTRATIVE EXAMPLES DEMONSTRATING THE UTILITY OF GLOBAL AND LOCAL FILTERING ON READ-ACROSS PREDICTIONS USING 5 CLOSEST ANALOGS

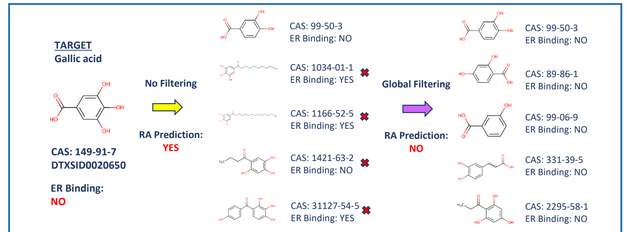


Figure 4: Predicting ER binding for Gallic Acid (non-binder) using analogs without filtering results in a read-across prediction: binder. Global filtering discards 4 analogs and selects new ones that meet the filtering criteria resulting in a read-across prediction: non-binder.

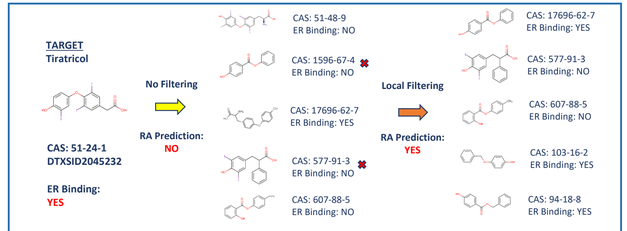


Figure 5: Predicting ER binding for Tiratricol (binder) using analogs without filtering results in a read-across prediction: non-binder. Local filtering discards 2 analogs and selects new ones that meet the filtering criteria resulting in a read-across prediction: binder.

Future Steps

Based on the complex interaction between the R-groups and their properties, and physchem properties of the chemical and ER binding, future research will focus on employing machine learning techniques to identify properties that are most relevant to these interactions.