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

Prachi Pradeep | pradeep.prachi@epa.gov | ORCID iD: 0002-9219-4249 | 919-541-5150

www.epa.gov

INTRODUCTION

Read-across is a data gap filling technique widely used within category and analog approaches to predict a biological property for a data-poor (target) chemical using known information from similar (source analog) chemical(s). Potential source analogs are typically identified based on structural similarity. Although much guidance has been published for read-across, practical principles for the identification and evaluation of the scientific validity of source analogs remains lacking.

This case study explores how well 3 structure descriptor sets (Pubchem, Chemotyper and MoSS) are able to identify analogs for read-across and predict Estrogen Receptor (ER) binding activity 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.

For each target chemical, analogs were selected using each descriptor set with two cut-offs: (1) Minimum Tanimoto similarity (range 0.1 - 0.9), and (2) Closest N analogs (range 1 - 10). Each target-analog pair was then evaluated for its agreement with measured ER binding and agonism. The analogs were then filtered using: (1) physchem properties of the analog, and (2) physchem properties of the ortho substituents (R-groups) of the analog. Both the analogs sets were subsequently filtered using number of literature sources as a marker for the quality of the experimental data. Finally, a majority vote prediction was made for each target phenol by reading-across from the closest *N* analogs. This case study presents an automated read-across approach which demonstrates: (1). that structural descriptors that are uninformed by properties driving the endpoint alone are insufficient for predicting the endpoint, regardless of the number of closest analogs or the Tanimoto similarity selected, and (2). quality of underlying experimental literature data profoundly influences the uncertainty of predictions.

OBJECTIVE

To investigate the utility of various structure descriptor methods for identification of analogs for read-across ER predictions and to assess the improvement in uncertainty of predictions by utilizing physchem properties, data quality measures and R-group properties for filtering of relevant analogs to ascertain better prediction of ER activity for hindered phenols.

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 read-across ER binding prediction was done for each target hindered phenol

Analog Selection Method

Descriptor Approach	Basis
Pubchem (P)	881 bits fingerprints
MoSS MCSS (M)	Size of most common substructure
Chemotyper (C)	Chemical substructures fingerprint with pre-defined chemotypes

Underlying basis for each of the three chemical descriptor approaches

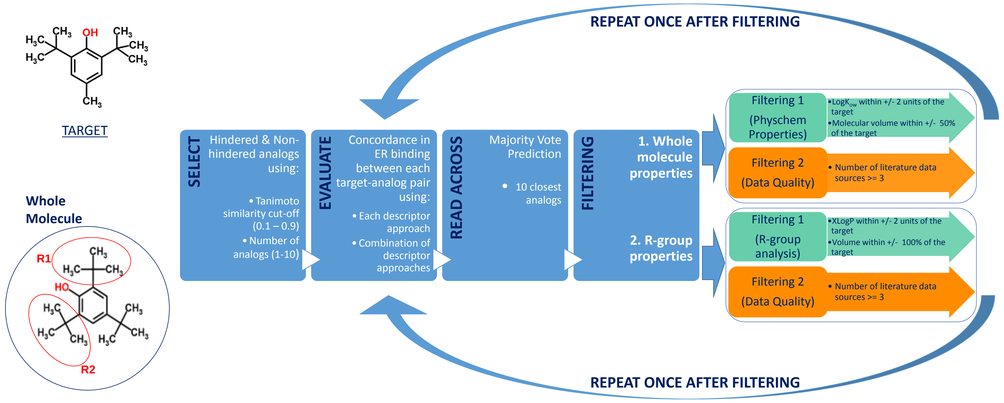
DATASET

Target: 462 hindered phenols
Inventory of Source Analogs: 719 phenols

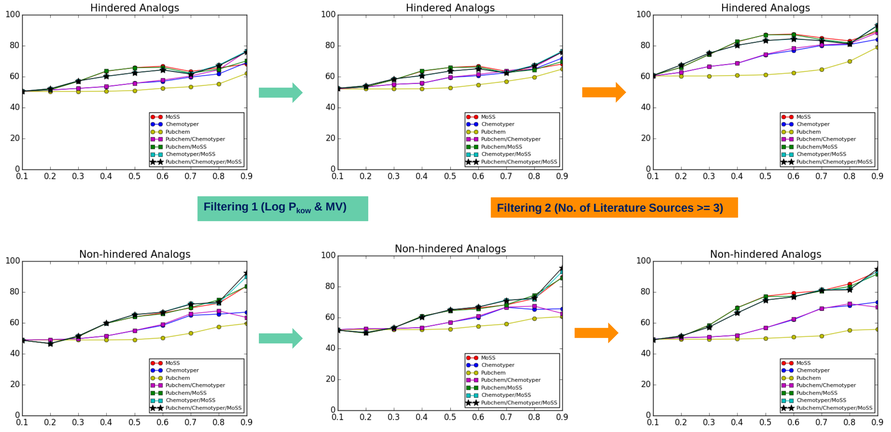
Number / Assay	Hindered		Non-hindered	
	Total	Actives	Total	Actives
Binding	462	207	257	155
Agonist	396	45	204	96
Antagonist	360	46	169	14

Distribution of hindered and non-hindered phenols with data in the source analog dataset

METHODS



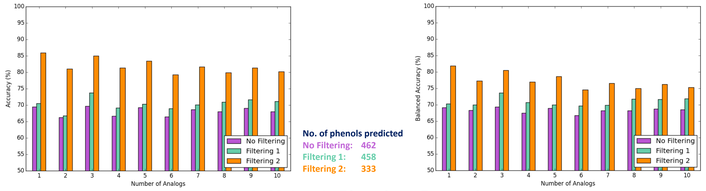
ANALYSIS



Disclaimer: The views expressed are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

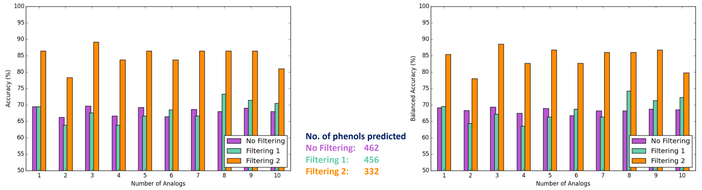
RESULTS

Read-across Accuracy and Balanced Accuracy (*E.g.* PubChem)



No. of phenols predicted
No Filtering: 462
Filtering 1: 458
Filtering 2: 333

Whole Molecule Physchem Properties



No. of phenols predicted
No Filtering: 462
Filtering 1: 456
Filtering 2: 332

R-group Analysis

CONCLUSIONS

- Analysis of analogs using a similarity cut-off (0.1-0.9) indicates that the concordance in estrogenic activity rises with increasing similarity. At a cut-off of 0.9, there is a marked increase in concordance (>80%). However, none of the descriptor approaches are distinctly better than the others in selecting analogs.
- Selecting hindered versus non-hindered phenols as analogs does not result in a significant improvement in concordance in estrogenic activity.
- Validity of analogs and the read-across prediction improves significantly after filtering of analogs based on bounds on endpoint related physchem properties, R-group properties and accounting for data quality.

The read-across predictions reveal that:

- PubChem and Chemotyper descriptors are superior to MoSS for ER activity.
- Filtering of analogs significantly increases the prediction accuracy. (E.g. the prediction accuracy using 3 closest analogs from PubChem is 70%. This increases to 74% when filtering by physchem properties, and 89% when data quality is accounted for.)

Future Directions:

- We see a complex interaction between the R-groups and their properties, and physchem properties of the whole molecule and probability of ER binding. Future research will focus on understanding these interactions, for instance using principal component analysis.

Identification of relevant and valid analogs for read-across prediction for any endpoint is not trivial. Using structural descriptors alone does not assure good performance. Addressing the key sources of uncertainty in read-across such as the quality of underlying experimental data and characterization of endpoint relevant properties lead to greatly improved read-across predictions.

U.S. Environmental Protection Agency
Office of Research and Development