Computational Toxicology 4 (2017) 22–30

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

A systematic evaluation of analogs and automated read-across prediction of estrogenicity: A case study using hindered phenols

Prachi Pradeep[a,b,*], Kamel Mansouri[a,b,1], Grace Patlewicz[b], Richard Judson[b]

[a] Oak Ridge Institute for Science and Education, Oak Ridge, TN, United States
[b] National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, United States

ARTICLE INFO

Keywords:
Read-across
Analog identification
Analog evaluation
Quantitative uncertainty analysis
Estrogen receptor (ER) binding

ABSTRACT

Read-across is an important data gap filling technique used within category and analog approaches for regulatory hazard identification and risk assessment. Although much technical guidance is available that describes how to develop category/analog approaches, practical principles to evaluate and substantiate analog validity (suitability) are still lacking. This case study uses hindered phenols as an example chemical class to determine: (1) the capability of three structure fingerprint/descriptor methods (PubChem, ToxPrints and MoSS MCSS) to identify analogs for read-across to predict Estrogen Receptor (ER) binding activity and, (2) the utility of data confidence measures, physicochemical properties, and chemical R-group properties as filters to improve ER binding predictions. The training dataset comprised 462 hindered phenols and 257 non-hindered phenols. For each chemical of interest (target), source analogs were identified from two datasets (hindered and non-hindered phenols) that had been characterized by a fingerprint/descriptor method and by two cut-offs: (1) minimum similarity distance (range: 0.1–0.9) and, (2) *N* closest analogs (range: 1–10). Analogs were then filtered using: (1) physicochemical properties of the phenol (termed global filtering) and, (2) physicochemical properties of the R-groups neighboring the active hydroxyl group (termed local filtering). A read-across prediction was made for each target chemical on the basis of a majority vote of the *N* closest analogs. The results demonstrate that: (1) concordance in ER activity increases with structural similarity, regardless of the structure fingerprint/descriptor method, (2) increased data confidence significantly improves read-across predictions, and (3) filtering analogs using global and local properties can help identify more suitable analogs. This case study illustrates that the quality of the underlying experimental data and use of endpoint relevant chemical descriptors to evaluate source analogs are critical to achieving robust read-across predictions.

Introduction

Read-across is a well-established data-gap filling technique used within category and analog approaches for regulatory hazard identification and risk assessment [1]. In the read-across approach, endpoint information for one or more chemicals (source analogs) are used to predict the same endpoint for another chemical (target), which is considered "similar" (usually on the basis of structural similarity) [1–3]. There are a number of steps in the development of a category or analog approach. Slight variations of the exact number and name of these steps depends on the technical guidance and publication used [1,4–6]. However, the two critical steps in the process are analog identification and analog evaluation [7,8]. Analog identification is the process of searching for source analogs similar to the target chemical. Source analogs are usually identified based on structural similarity, using fingerprints that encode chemical information based on the presence or absence of certain structural features [5,9]. A similarity index such as the Jaccard distance (Tanimoto index) [10] is then used as a threshold to limit the number of source analogs retrieved. Many web-based tools that permit structure searching include an algorithm to search for structurally similar chemicals in this manner. Common web based tools include ChemID plus [11], and ChemSpider [12]. This type of structural search tends to be general in scope, in the sense that no assumptions are made to limit the analog search on the basis of properties or parameters that might be pertinent to a specific endpoint. On the other hand, a search for source analogs informed by parameters relevant to the endpoint of interest would rely on descriptors which could affect chemical bioavailability and reactivity, such as physicochemical properties (e.g., LogP, molecular volume), electronic properties (e.g., energy of the lowest unoccupied orbital (eLUMO), energy of

[*] Corresponding author at: Oak Ridge Institute for Science and Education, Oak Ridge, TN, United States.
E-mail address: pradeep.prachi@epa.gov (P. Pradeep).
[1] Current affiliation: ScitoVation LLC, 6 Davis Dr, Research Triangle Park, NC 27709, United States.

http://dx.doi.org/10.1016/j.comtox.2017.09.001
Received 7 June 2017; Received in revised form 24 July 2017; Accepted 7 September 2017
Available online 08 September 2017

the highest occupied orbital (eHOMO) [5]. The next step, analog evaluation, gathers associated property and effect information for the source analogs to determine their relevance and suitability for the endpoint of interest. Many structure fingerprint and descriptor methods are available (free or commercial), each of which capture different aspects of chemical structure that are potentially relevant to different endpoints.

Despite available guidance [3,4,7] for analog/category approaches to read-across, guiding principles to evaluate analog validity for specific endpoints remains lacking [4,8]. The similarity rationale underpinning source analog selection, as well as the quantity and quality of experimental data associated with the selected analogs, are important sources of uncertainty in read-across [4,13]. Consequently, even though read-across is conceptually accepted by both regulatory agencies and industry, difficulties remain in the consistent application of read-across approaches in practice, which in turn has limited their acceptance for regulatory decisions [3,14]. Efforts have been made to standardize and characterize a framework for documenting read-across justifications to help increase consistency and promote regulatory acceptance of read-across predictions [4,13,15]. Although several read-across studies have been published recently [16–18], successful examples are still lacking [14].

To establish improved and reproducible read-across predictions, this case study undertook a systematic analysis of analogs for read across predictions of *in vitro* ER binding. Hindered phenols were selected as an example chemical class. Hindered phenols are defined as phenols that contain one or more bulky functional groups ortho to the phenolic hydroxyl group, e.g., 3-chloro-4-hydroxybenzoic acid. Phenols, in general, are known to mimic the activity of estrogen and possess estrogenic activity resulting in the possibility of endocrine disruption [19,20]. Endocrine disruption can lead to a wide range of health disorders in humans, including reproductive and developmental toxicity [21,22]. Phenols can interact with the estrogen receptor (ER) due to the presence of the phenol hydroxyl group, which aids in binding with ER. Hindered phenols are expected to be less potent ER binders than non-hindered phenols because their bulky functional groups block the hydroxyl group-protein interaction [23].

An automated approach was developed for this case study to: (1) identify and evaluate the suitability (validity) of source analogs; and (2) evaluate and assess uncertainty due to confidence in data and analog suitability in read-across predictions. Specifically, the case study presents an analysis of the ability of three structure fingerprint/descriptor methods to identify source analogs to read-across ER binding, and the use of data confidence measures, physicochemical properties, and chemical substituent functional (R) group physicochemical properties to evaluate the validity of the source analogs identified.

Methods

Dataset

The dataset of phenols used in this study was extracted from the prediction dataset constructed as part of CERAPP, the Collaborative Estrogen Receptor Activity Prediction Project [24]. This CERAPP prediction dataset (herein referred to as the source dataset) contained literature-derived curated data from a number of overlapping sources including Tox21 [25–29], U.S. FDA Estrogenic Activity Database (EADB) [30], METI (Ministry of Economy, Trade and Industry, Japan) database [31], and ChEMBL [32] for over 32,000 chemical structures. Each chemical in the CERAPP source dataset had been assigned a literature data source count such that there was $< 20\%$ disagreement among different sources. For instance, if there were 4 independent publications of ER activity for a chemical, all 4 sources had to agree (i.e., ER binder or non-binder) for the chemical to be included in the source dataset. On the other hand, if there were 5 published reports for a particular chemical, the chemical would still be included in the source dataset if one reference disagreed on ER binding activity with the majority consensus. The majority ER activity outcome from all the sources was taken as the final outcome, 1 or 0 representing ER binder and non-binder, respectively. The literature data source count from CERAPP was used as a surrogate for data confidence in this study. The expectation was that the more consistent the literature reports were of ER activity (binder or non-binder), the more likely the activity could be relied upon to be reproduced in a subsequent experiment. A custom KNIME workflow (version 2.11.3) [33] was developed to extract phenols from the larger CERAPP chemical library and to categorize them as being hindered or not, based on the presence or absence of bulky groups at the ortho position. The final dataset used in this study comprised 719 phenols with 462 hindered phenols (207 ER binders) and 257 non-hindered phenols (155 ER binders).

Chemical descriptors

One study aim was to analyze the ability of different structure descriptor approaches to identify source analogs for hindered phenols, and evaluate the adequacy of the analogs for ER read-across predictions. This enabled a baseline performance assessment to be made for the preliminary analogs identified. While a myriad of fingerprints/descriptors can be computed for chemical properties (structure, physicochemical, electronic), there are no published or systematic guidelines for evaluating the suitability of one descriptor type versus another for a specific endpoint.

Three common structure-based fingerprint/descriptors sets (PubChem [34], ToxPrints [35], and MoSS MCSS [36]) were used in this study. PubChem fingerprints are 881 bits long where each bit represents the presence or absence of a specific substructure. The substructure categories spanned by a PubChem fingerprint include hierarchical element counts, rings in a canonical extended smallest set of smallest rings, ring set, simple atom pairs, simple atom nearest neighbors, detailed atom neighborhoods, simple SMARTS patterns, and complex SMARTS patterns. The PubChem fingerprints were generated in KNIME analytics platform (version 2.11.3) [33]. ToxPrint chemotypes (or ToxPrints) comprise 729 uniquely defined chemical features (https://toxprint.org) coded in XML-based Chemical Subgraphs and Reactions Markup Language (CSRML) [35]. The ToxPrints features were specifically designed to provide a broad coverage of inventories consisting of environmental and industrial chemicals including pesticides, cosmetics ingredients, food additives and drugs. The fingerprints represent a wide range of substructures comprising atoms, bonds, chains, groups and ring elements. The ToxPrints were generated within the publically available Chemotyper application (version1.0.r12976, https://chemotyper.org). MoSS is a substructure miner algorithm implemented in KNIME analytics platform (version 2.11.3) [33] that calculates the size of the maximum common substructure (MCSS) between two chemicals. The MCSS of two chemicals is the largest possible substructure that is present in both chemical structures; more similar the chemicals have larger MCSS sizes. The Jaccard distance (Tanimoto index) was used to calculate pairwise similarity indices for the all phenols in the datasets as characterized by the PubChem and ToxPrints descriptor sets. The similarity index ranges from 0-1, where 0 indicates least similar (dissimilar) and 1 indicates most similar (mostly chemical similarity by itself). These pairwise similarities were summarized in a similarity matrix. The similarity matrix is calculated using a component called a distance matrix in KNIME. For the third descriptor set, the MCSS itself was taken as the similarity index.

Read-across analysis workflow

Fig. 1 summarizes the four steps of the read-across analysis workflow that was followed in this study. First, a set of structurally related analogs were identified using each of the three descriptor sets to determine the baseline performance of ER read-across predictions for the