Short Communication

Generalized Read-Across (GenRA): A Workflow Implemented into the EPA CompTox Chemicals Dashboard

George Helman[1,2], Imran Shah[2], Antony J. Williams[2], Jeff Edwards[2], Jeremy Dunne[2] and Grace Patlewicz[2]

[1]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA; [2]National Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency, Research Triangle Park (RTP), NC, USA

Abstract

Generalized Read-Across (GenRA) is a data driven approach that makes read-across predictions on the basis of a similarity weighted activity of source analogues (nearest neighbors). GenRA has been described in more detail in the literature (Shah et al., 2016; Helman et al., 2018). Here we present its implementation within the EPA's CompTox Chemicals Dashboard to provide public access to a GenRA module structured as a read-across workflow. GenRA assists researchers in identifying source analogues, evaluating their validity and making predictions of *in vivo* toxicity effects for a target substance. Predictions are presented as binary outcomes reflecting the presence or absence of toxicity together with quantitative measures of uncertainty. The approach allows users to identify analogues in different ways, quickly assess the availability of relevant *in vivo* data for those analogues, and visualize these in a data matrix to evaluate the consistency and concordance of the available experimental data for those analogues before making a GenRA prediction. Predictions can be exported into a tab-separated value (TSV) or Excel file for additional review and analysis (e.g., doses of analogues associated with production of toxic effects). GenRA offers a new capability of making reproducible read-across predictions in an easy-to-use interface.

1 Introduction

Given the thousands of data-poor or toxicologically uncharacterized chemicals in commerce, read-across has proved to be a convenient and efficient data gap filling technique that can be used within analogue and category approaches for many different regulatory purposes. Read-across represents the application of data from a source chemical(s) for a particular property or effect to predict the same property or effect for the target chemical (the chemical of interest) (OECD, 2014). Read-across is traditionally anchored with conventional *in vivo* and *in vitro* data, though concerted efforts are starting to be made to exploit high throughput (HT) and high content (HC) screening data as a means of substantiating biological similarity (Zhu et al., 2016; Shah et al., 2016). Some of these efforts are anchoring such data to key events within adverse outcome pathways (AOPs) (Schultz and Cronin, 2017).

Here we present the web-based implementation of Generalized Read-across (GenRA), a data-driven approach that makes reproducible read-across predictions of toxicity outcomes from *in vivo* studies (Shah et al., 2016). The read-across prediction is a similarity weighted activity of source analogues (nearest neighbors) based on chemistry and/or bioactivity descriptors. The approach is a generalization of the Chemical Biological Read-Across (CBRA) approach published by Low et al. (2013). GenRA has been described in more detail in the literature (Shah et al., 2016; Helman et al., 2018). Here we outline the principles of the approach and its workflow implementation in the EPA CompTox Chemicals Dashboard (Williams et al., 2017;[1]).

2 Materials and methods

The GenRA framework has been implemented using a classical three tier architecture, which is seamlessly embedded in the EPA CompTox Chemicals Dashboard, and includes: 1) a web-based presentation tier; 2) an application tier based on representational

[1] https://comptox.epa.gov/dashboard

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Received November 29, 2018; Accepted January 30, 2019;
Epub February 4, 2019; © The Authors, 2019.

ALTEX 36(3), 462-465. doi:10.14573/altex.1811292

Correspondence: Grace Patlewicz, PhD; National Center for Computational Toxicology (NCCT), US Environmental Protection Agency
109 TW Alexander Dr, Research Triangle Park (RTP), NC 27711, USA
(patlewicz.grace@epa.gov)

state transfer (REST) web services; and 3) a data tier for storing large-scale chemical, bioactivity, and toxicity data for thousands of chemicals. The presentation tier of GenRA is implemented using Vue[2] and each step in the workflow is designed as a self-contained component. Each component intuitively captures the key tasks that must be performed in the workflow via a combination of inputs (i.e., buttons, input items, etc.) and an interactive graphical output. All graphical outputs of the individual components are implemented as scalable vector graphic (SVG) elements with context sensitive help information and/or interaction capabilities. The presentation layer components in GenRA perform their specific tasks by obtaining information about chemicals, analogues, bioactivity and toxicity from the application tier. The application tier is implemented in Python[3] using the Flask[4] microservices framework, which is deployed using Apache/wsgi[5]. The data tier is implemented using MongoDB[6], which is a document-oriented NoSQL database. Information about chemicals, bioactivity and toxicity are stored as separate MongoDB collections to facilitate the efficient implementation of GenRA algorithms. Chemical structure data were obtained from the Distributed Structure Searchable Toxicity (DSSTox) database (originally extracted April 2017 but updated continuously) (Richard et al., 2016;[7]) whereas chemical descriptors, comprising Morgan fingerprints (Rogers and Hahn, 2010) and topological torsion descriptors (Nilakantan et al., 1987), were generated using RDKit[8]. ToxPrint chemotypes were generated using the AM-MN Chemotyper for command line operation (Yang et al., 2015;[9]).

The bioactivity high throughput screening (HTS) data were obtained from the ToxCast[10] and Tox21[11] programs. The *in vivo* toxicity data was obtained from ToxRefDB v.1.0[12].

Bioactivity descriptors (denoted biology or bio) comprised hit calls (active (1) and inactive (0)) from 820 ToxCast HTS assays. The 820 bioactivity descriptors were converted into fingerprints that are used singly (chm or bio to denote either chemical or bioactivity descriptors) to predict up to 129 toxicity outcomes from 10 different study types from ToxRefDB v1.0. The study types are namely acute (ACU), subacute (Sub), subchronic (SAC), neurotoxicity (NEU), developmental neurotoxicity (DNT), developmental toxicity (DEV), reproductive toxicity (REP), and multigenerational toxicity (MGR). A final category of other (OTH) is for any study not fitting any of the previously mentioned study types.

3 Results and discussion

There are several steps in the development of a category or analogue approach (Patlewicz et al., 2017, 2018). The seven key steps in the workflow are as follows:

1. Decision context
2. Data gap analysis
3. Overarching similarity rationale
4. Analogue identification
5. Analogue evaluation
6. Data gap filling
7. Uncertainty assessment

In the GenRA implementation, the steps have been addressed as shown in Figure 1 (Helman et al., 2018).

The starting point for GenRA relies on identifying the chemical of interest (target chemical) by performing a "basic" search within the EPA CompTox Dashboard. The outcome of a search gives rise to a "chemical details" landing page with a number of selectable tabs and sub-tabs to the left of the screen (Fig. S1[13]). One of the tabs navigates to the GenRA module.

Once GenRA is selected, a grid like display is presented with an indicator at the top of the page that reflects the relevant step in the read-across workflow. Users can navigate between steps by clicking on the indicator bar (Fig. S2[13]).

The starting grid display only has the first window unobscured. This grid window shows the neighborhood of source analogues that surround the target substance which appears in the center of a radial plot (Fig. S3[13]). Starting from 12-o'clock on the plot, analogues are ordered in decreasing order of similarity as calculated by the Jaccard index[14] (which ranges from 0 to 1, where 0 denotes dissimilar and 1 denotes identical). This radial plot represents the analogue identification and evaluation steps of the workflow. By default, 10 analogues are shown which are based on Morgan chemical fingerprints. The view can be updated by choosing a different fingerprint type and by changing the number of analogues. A minimum of 5 analogues and maximum of 10 analogues can be selected. Analogues are automatically filtered by the availability of *in vivo* toxicity data as taken from ToxRefDB v1.0. This is to ensure that analogues identified are helpful in a read-across prediction. Hovering over any of the source analogue depictions in the radial plot reveals the numerical pairwise similarity between the target and that of the analogue. If the user wishes to

[2] http://vuejs.org
[3] https://www.python.org/
[4] http://flask.pocoo.org/
[5] https://www.python.org/dev/peps/pep-3333/
[6] https://www.mongodb.com/
[7] https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database
[8] http://www.rdkit.org/
[9] https://chemotyper.org
[10] https://www.epa.gov/chemical-research/toxicity-forecasting
[11] https://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21
[12] ftp://newftp.epa.gov/COMPTOX/High_Throughput_Screening_Data/Animal_Tox_Data/readme_toxrefdb_20141106.pdf (accessed 31.01.2019).
[13] doi:10.14573/altex.1811292s
[14] https://en.wikipedia.org/wiki/Jaccard_index

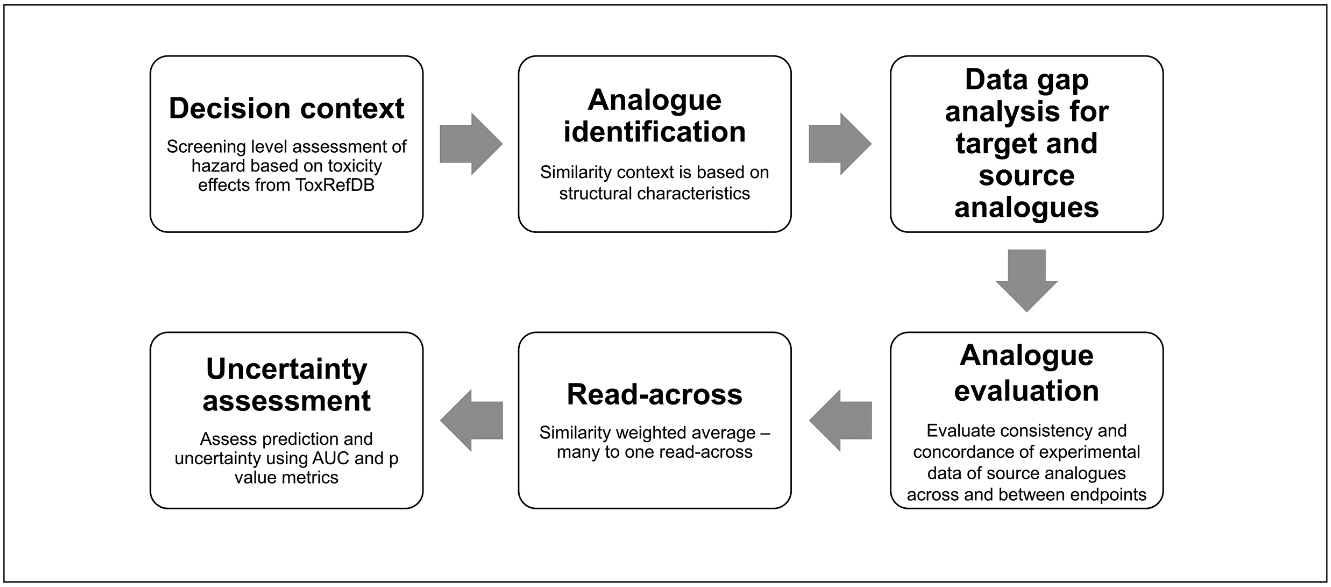
Decision context
Screening level assessment of hazard based on toxicity effects from ToxRefDB
Analogue identification
Similarity context is based on structural characteristics
Data gap analysis for target and source analogues
Analogue evaluation
Evaluate consistency and concordance of experimental data of source analogues across and between endpoints
Read-across
Similarity weighted average – many to one read-across
Uncertainty assessment
Assess prediction and uncertainty using AUC and p value metrics

Fig. 1: Category/analogue workflow and GenRA
This figure has been reproduced with permission from Helman et al. (2018).

conduct a GenRA analysis for a different source analogue, or wishes to view the Chemical Results page, clicking on the structure depiction in the radial plot will open a new browser tab with the respective chemical details page of that analogue in the Dashboard. Once a user is satisfied with the analogues identified, the Next button needs to be clicked to proceed to the next step of the workflow – Data gap analysis (denoted as Step Two: Data Gap Analysis and Generate Data Matrix in the interface).

At this point, the next two grid views become unobscured and the workflow indicator changes to “Data Gap Analysis & Generate Data Matrix”. The first of these grid views is denoted as “Summary Data Gap Analysis” (Fig. S4[13]). This view is intended to provide a landscape of the quantity of data records for the target substance and its source analogues with respect to different data streams listed earlier – ToxCast, Tox21, Chemotypes and ToxRefDB. The number of records is marked in the colored boxes and reflected in the color itself – the black box indicates the greatest number of records whereas a yellow box indicates fewer records. Colors are automatically assigned by the underlying number of records. The summary is to provide a rapid perspective of how feasible a read-across might be based on the quantity of data for the source analogues.

The second grid view (Fig. S5[13]) reflects ToxRef as a group by Tox Fingerprint. In this case, the data view shows the *in vivo* toxicity effect records as represented by this toxicity fingerprint. A black colored box in the grid view denotes the presence of a record for a particular toxicological effect. The utility of the grid view is to help a user gain a perspective of what data gaps exist for the source analogues relative to the target substance, and which effects might be reasonably predicted by those analogues. The entire matrix can be browsed using the scroll bar. A user might choose to focus on a subset of effects with the knowledge that the identified analogues will be helpful in that regard as data are available or alternatively that the analogue set is missing the data for the toxicity effects of most interest.

Whilst the summary views are helpful to gain a brief perspective of how much data are available for the target and source analogues, they do not provide any information on their potency (e.g., lowest effect limit (LEL) in mg/kg-day, etc.) or hazard profile. To evaluate this type of information, the “Generate Data Matrix” button is clicked to move to the next step of the workflow “Run GenRA Prediction”. At this point, the final grid becomes unobscured to reveal a matrix view of the target and source analogues. The initial part of the assessment here addresses the “Analogue Evaluation” step since the user can evaluate the consistency and concordance of the analogues, relative to their experimental data, in terms of the presence or absence of toxicity effects. Presence and absence is reflected by the colors of the boxes in the data matrix: red for the presence of toxicity effects, blue for the absence of toxicity effects, and grey for no data. Hovering over any box reveals a tool tip indicating no data, or no effects for grey and blue colors, respectively, whereas the doses at which toxicity effects were reported are shown for red boxes. The data matrix view, using the same color codes (Fig. S6[13]), provides the user with an informative perspective of the consistency and concordance of the available data across the analogues and between the endpoints. Users can filter the effects of interest using the filter window, select the threshold for the number of positives and negatives within the analogue set, and alter the view so that the similarity index is used to shape the size of the data matrix boxes (Fig. S7[13]). The data matrix is ordered by the target substance in the first column, followed by the source analogues in order of decreasing similarity.

The full extent of toxicity effects can be browsed by using the scroll bar to the right of the screen. Users can also elect to deselect a source analogue from further consideration by clicking on the tick symbol by the similarity index value at the top of the column. Once the user has selected the desired source analogues, the "Run Read-across" button is clicked to derive the GenRA predictions. There is a short time lag whilst the calculations are performed. Although the actual predictions are computed rapidly (given the simplicity of the similarity-weighted activity algorithm), the calculation of the associated performance metrics, which provide the uncertainty assessment characteristics, takes a little longer to process. After the GenRA predictions are generated, the first column in the data matrix view (the column for the target substance) is updated with colors of differing opacity (Fig. S8[13]). The colors are still red or blue, but the degree of opacity denotes the confidence associated with any prediction. The darker the color, the less confident the prediction (since opacity is scaled by the lower p-value). The confidence is measured by 2 characteristics, the area under the curve (AUC) of a ROC (receiver operating characteristic) and the p-value (see Shah et al., 2016 for further details). The higher the AUC and the lower the p-value, the more confident the prediction. The current GenRA implementation is focused on structural or bioactivity predictions; other contexts of similarity (such as metabolism) that are pertinent in traditional read-across, will be the subject of future work.

The final step of the workflow involves exporting the predictions generated using the download options. File types include TSV and Excel files which can be exported and mirror the view presented in the application data matrix (Fig. S9[13]). Currently, any subsequent analysis requires the end user to exploit other data analysis tools. In future releases, options to sort and rank order predictions and aggregate by toxicity type will be available, as well as other possibilities to identify and evaluate analogues.

GenRA provides the user with a standardized workflow interface to enable reproducible data driven read-across predictions of *in vivo* toxicity effects based on chemical/or bioactivity fingerprints. The predictions are binary outcomes of presence or absence of toxicity with quantitative measures of uncertainty as expressed by the AUC and p-values. In addition, exported results include potency values of analogues that can be further analyzed. The inclusion of the GenRA module into the CompTox Chemicals Dashboard has provided community access to generate read-across predictions in a highly intuitive manner. The GenRA workflow and user documentation are available on the EPA CompTox Chemicals Dashboard website[15].

References

Helman, G., Shah, I. and Patlewicz, G. (2018). Extending the generalised read-across approach (GenRA): A systematic analysis of the impact of physicochemical property information on read-across performance. *Comput Toxicol 8*, 34-50. doi:10.1016/j.comtox.2018.07.001

Low, Y., Sedykh, A., Fourches, D. et al. (2013). Integrative chemical-biological read-across approach for chemical hazard classification. *Chem Res Toxicol 26*, 1199-1208. doi:10.1021/tx400110f

Nilakantan, R., Bauman, N., Dixon, J. S. and Venkataraghavan, R. (1987). Topological torsion – A new molecular descriptor for SAR applications – Comparison with other descriptors. *J Chem Inf Comput Sci 27*, 82-85. doi:10.1021/ci00054a008

OECD – Organisation for Economic Co-operation and Development (2014). Guidance on Grouping of Chemicals, Second Edition. *OECD Series on Testing and Assessment No. 194*. OECD Publishing, Paris, France. doi:10.1787/9789264274679-en

Patlewicz, G., Helman, G., Pradeep, P. and Shah, I. (2017). Navigating through the minefield of read-across tools: A review of in silico tools for grouping. *Comput Toxicol 3*, 1-18. doi:10.1016/j.comtox.2017.05.003

Patlewicz, G., Cronin, M. T. D., Helman, G. et al. (2018). Navigating through the minefield of read-across frameworks: A commentary perspective. *Comput Toxicol 6*, 39-54. doi:10.1016/j.comtox.2018.04.002

Richard, A. M., Judson, R. S., Houck. K. A. et al. (2016). The ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chem Res Toxicol 29*, 1225-1251. doi:10.1021/acs.chemrestox.6b00135

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *J Chem Inf Model 50*, 742-754. doi:10.1021/ci100050t

Schultz, T. W. and Cronin, M. T. D. (2017). Lessons learned from read-across case studies for repeated-dose toxicity. *Regul Toxicol Pharmacol 88*, 185-191. doi:10.1016/j.yrtph.2017.06.011

Shah, I., Liu, J., Judson, R. S. et al. (2016). Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. *Regul Toxicol Pharmacol 79*, 12-24. doi:10.1016/j.yrtph.2016.05.008

Williams, A. J., Grulke, C. M., Edwards, J. et al. (2017). The CompTox chemistry dashboard – A Community data resource for environmental chemistry. *J Cheminform 9*, 61. doi:10.1186/s13321-017-0247-6

Yang, C., Tarkhov, A., Marusczyk, J. et al. (2015). New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J Chem Inf Model 55*, 510-528. doi:10.1021/ci500667v

Zhu, H., Bouhifd, M., Donley, E. et al. (2016). Supporting read-across using biological data. *ALTEX 33*, 167-182. doi:10.14573/altex.1601252

Conflict of interest

The authors declare that they have no conflict of interests.

Acknowledgements

The authors wish to thank the NCCT software development team for their dedication and commitment to the delivery of the GenRA module and ongoing development of the CompTox Chemicals Dashboard .

[15] https://comptox.epa.gov/dashboard