Bioinformatics, 38(7), 2022, 2066–2069
https://doi.org/10.1093/bioinformatics/btac045
Advance Access Publication Date: 3 February 2022
Applications Note

Gene expression

EDTox: an R Shiny application to predict the endocrine disruption potential of compounds

Amirhossein Sakhteman, Arindam Ghosh and Vittorio Fortino *

School of Medicine, Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland

*To whom correspondence should be addressed.
Associate Editor: Anthony Mathelier

Received on September 3, 2021; revised on January 11, 2022; editorial decision on January 13, 2022

Abstract

Purpose: Endocrine disruptors are a rising concern due to the wide array of health issues that it can cause. Although there are tools for mode of action (MoA)-based prediction of endocrine disruption (e.g. QSAR Toolbox and iSafeRat), none of them is based on toxicogenomics data. Here, we present EDTox, an R Shiny application enabling users to explore and use a computational method that we have recently published to identify and prioritize endocrine disrupting (ED) chemicals based on toxicogenomic data. The EDTox pipeline utilizes previously trained toxicogenomic-driven classifiers to make predictions on new untested compounds by using their molecular initiating events. Furthermore, the proposed R Shiny app allows users to extend the prediction systems by training and adding new classifiers based on new available toxicogenomic data. This functionality helps users to explore the ED potential of chemicals in new, untested exposure scenarios.

Availability and implementation: This tool is available as web application (www.edtox.fi) and stand-alone software on GitHub and Zenodo (https://doi.org/10.5281/zenodo.5817093).

Contact: vittorio.fortino@uef.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Endocrine disruptors are chemicals that have the potential to alter the functioning of the endocrine systems of an organism. They do this by mimicking, blocking or interfering with the hormones and hence can result in a wide array of health issues. Here, we present EDTox, an R Shiny application that helps non-expert R users to utilize our previously published method (Sakhteman *et al.*, 2021) that allows identifying the putative mode of action (MoA) of compounds with suspected endocrine disrupting (ED) activity, and classifying untested compounds as EDC or not based on their known interaction with genes or the molecular initiating events (MIEs). The set of genes interacting with compounds are selected from the Comparative Toxicogenomics Database (CTD; Davis *et al.*, 2021). These associations are based on five interaction types: reaction, binding, activity, expression and metabolic processing. The highlight of this application is its ability to compile an EDC-class probability score which indicates whether a given compound is strongly related to endocrine disruption for a specific exposure scenario (e.g. rat liver after 24 h of toxicant exposure). The EDTox application is divided into six tabs. Figure 1 graphically illustrates the implemented software modules.

***Home* tab:** The *Home* tab is the primary landing tab when the application is launched. It contains a brief introduction about the application and the background data used for building the EDTox pipeline. It also includes the files that are needed to implement the described case study.

***Summary* tab:** In our previous study, we had used data from large-scale toxicogenomics projects—LINCS , TG-GATEs (Igarashi *et al.*, 2015) and DrugMatrix (Ganter *et al.*, 2005)—to train and validate classifiers for predicting compounds with ED potential in different exposure scenarios (Sakhteman *et al.*, 2021). The *Summary* tab provides an overview of the accuracy of these pretrained classifiers, which can be used to make new prediction. Moreover, this tab can also be used to compare the accuracy of any new classifier trained using the *Toxicogenomics pipeline* tab of the application.

***Toxicogenomics pipeline* tab:** This is the main section where the EDTox pipeline is implemented to train a new classifier based on new available toxicogenomics signatures. The users here are required to submit three primary inputs: (i) a toxicogenomics data derived weighted undirected gene–gene coexpression network; (ii) a list of EDCs (as MeSH ID or CAS number) along with their MIEs (as entrez gene ID) and (iii) a list of negative controls along with their MIEs. To be noted here that the EDTox application does not include any module for construction of the network. Users are required to build the network separately using a tool of their choice for input with the EDTox application. The required format for the input files is described within Supplementary Files S1 and S3.

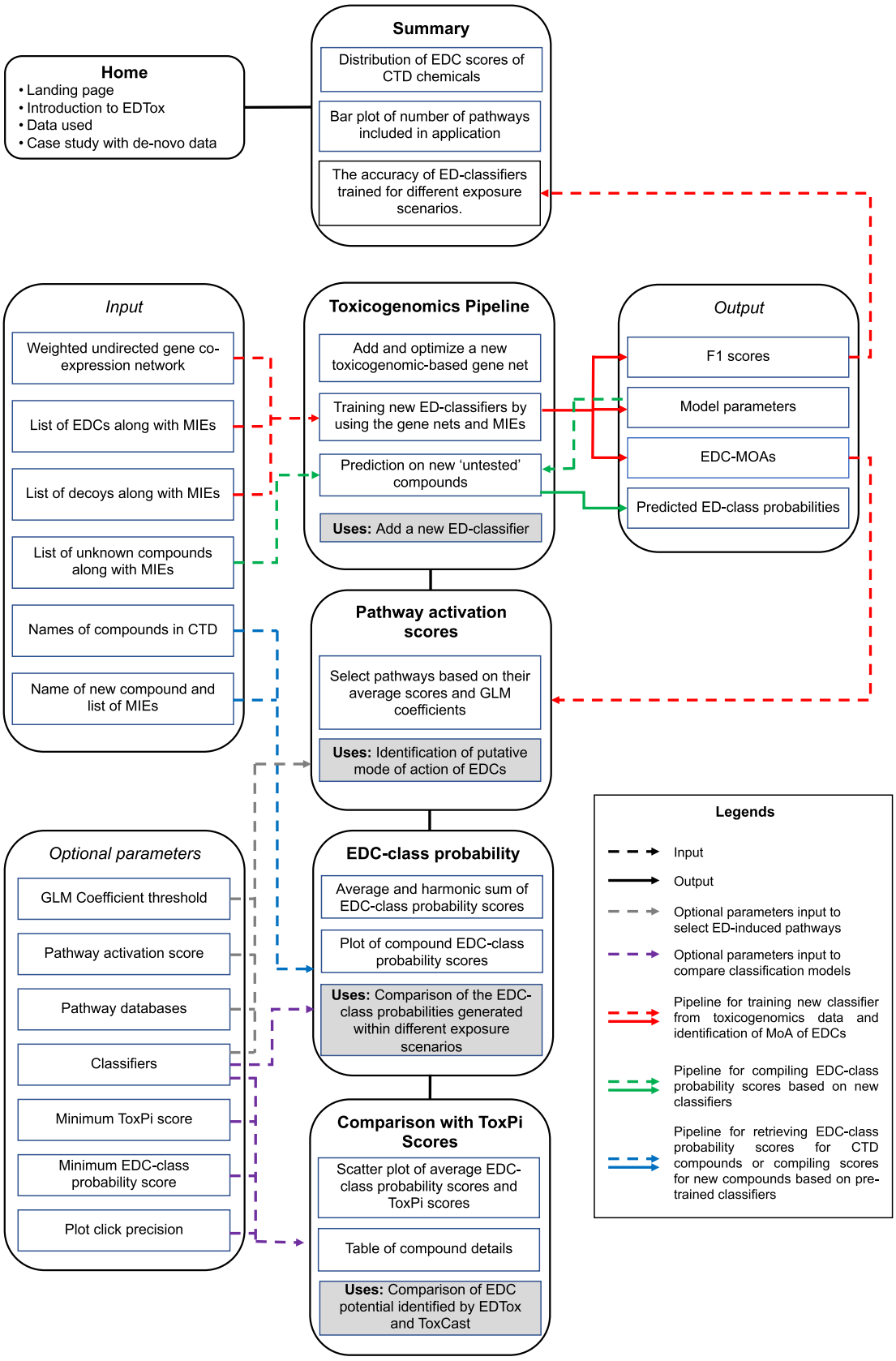

Fig. 1. Overview of the software modules provided by the EDTox platform. Each module includes a brief description of the implemented functionalities. The different functional relationships between the modules of EDTox are indicated by using different arrows and colors. Each color refers to a specific pipeline or a set of tunable parameters.

The EDTox pipeline for training the classifier utilizes general linear models (GLMs). This step is implemented within the *Training and validation of GLM-based classifier* module of this tab. The underlying pipeline implemented to train the classifier is described in further details in Sakhteman *et al.* (2021).

***Pathway activation scores* tab:** It helps identify the putative MoA of the EDCs. The use of GLM elastic net classifier in the EDTox pipeline allows to identify the features that are prominent in forming the basis of classification. The *Pathway activation scores* tab provides a systematic comparison of average pathway activation scores and GLM coefficients of pathways for different classifiers that can be used to identify the putative MoA of the EDCs. By exploring differences in activated pathways, users can study the specific MoA of compounds under different exposure scenarios.

***EDC-class probability scores* tab:** To retrieve previously computed EDC-class probability scores for compounds in CTD and compare them based on different classifiers of up to five compounds. The average and harmonic sum of the scores for each compound is also shown within this tab. The tab also allows the user an access to the pretrained classifiers to compile the EDC-class probability scores for compounds not in CTD or other untested compounds based on an input list of MIEs for the compound. MIEs (or the genes interacting with the untested compound) can be characterized by using *in silico* molecular docking or *in vitro* assay data (e.g. ToxCast), and as last resort, by *in vivo* testing.

***Comparison with ToxPi scores* tab:** The Toxicological Priority Index (ToxPi) score is one of the widely used parameters for prioritizing chemicals based on their toxicity (Filer *et al.*, 2014). The *Comparison with ToxPi scores* tab allows the user to compare the average of EDC scores of the CTD chemicals compiled from this application with the ED-based ToxPi scores from ToxCast. The details of each compound including links to CompTox database are rendered as a table on selecting a compound from the scatter plot.

3 Use cases

Making predictions for untested compounds: The user provides in input the name of an untested compound, which can correspond to an MeSH ID, CAS number or a generic name (if the compound is not included in CTD), along with a list of genes (e.g. 11 140, 8676, 7323) with which it interacts, within the *Compile EDC-class probability scores for new chemicals* module of the *EDC-class probability scores* tab of the application. The selected classifiers will then compute ED class scores indicating whether a given compound is strongly related to endocrine disruption for a specific exposure scenario (e.g. rat liver after 24 h of toxicant exposure). The tab also reports the average and harmonic sum of the scores within a table while displaying the scores for individual classifiers (Supplementary File S1 and Fig. S11).

Training new classifiers from *de novo* toxicogenomic data: The users can train new classifiers for prediction of EDCs using the *Toxicogenomics pipeline* tab. For example, starting from the gene expression data of HepaRG cells treated with 20 carcinogens for 24 h (S-DIXA-AN-012), the users are required to create a weighted gene–gene coexpression network as one of the primary inputs for training the classifier (R script for preparing the network included in Supplementary File S2). Apart from this, they are also required to upload lists of EDCs and negative controls and their MIEs or interacting genes that will be used as training set for training the classifier. Next, the users can opt to use the *Selection of optimal parameters for RWR-FGSEA* module to select the proportion of edges from the input network to be used for random walk with restart (RWR) and the number of top ranked genes from RWR to be considered for fast gene set enrichment analysis (FGSEA). The selected parameters (in our case, a combination of top 2% edges and top 1000 genes) (Supplementary File S2 and Fig. S1) can then be finally used to execute the RWR-FGSEA-GLM step whereby the selected gene network is used for compiling the pathway activation scores followed by training and evaluation of a GLM-based classifier. The completion of execution of the pipeline returns a plot of the F1 accuracy scores of the trained model over the *k*-fold cross validation (Supplementary File S2 and Fig. S2). The model parameters, its accuracy scores and the identified MOAs for the EDCs could be exported using the *Export* module. The saved F1 scores for this newly trained classifier could be compared to the pretrained classifier by uploading it in the *Summary* tab of the application.

Using *de novo* classifiers for compiling ED scores: The user exploits the *Prediction of new compounds* module to test previously trained classifiers and calculate the EDC-class probability scores and EDC-classes for untested compounds. Here, the model parameters obtained after training the classifier is needed to be uploaded along with list of unknown compounds (as MeSH ID or CAS number) and their MIEs (as entrez gene ID or gene symbols). The module returns a table containing the EDC-class probability scores and the EDC-class which can also be saved using the *Export* module. For our case study, we used a random set of 50 compounds from CTD along with their MIEs and predicted their class probabilities (Supplementary File S2 and Table S1).

Exploring the MoA of compounds with suspected ED activity: The EDTox application can be used for identification of the MOA of the EDCs. This is done using the bubble plot in the Pathway activation scores tab. For example, for the classifier trained using the gene expression data of HepaRG cells treated with carcinogens for 24 h, the user can set threshold values for the coefficients estimated for each pathway (e.g. 0.1) and the overall pathway activation score (e.g. 0.5), in order to compare the activated pathways from HepaRG model with other *in vitro*-based data layers (Supplementary File S2 and Fig. S4).

4 Conclusion

The EDTox R Shiny application offers a user-friendly graphical interface that allow non-expert R users to utilize large-scale toxicogenomics data for the identification and prioritization of ED compounds. EDTox also allows to explore the molecular pathways that were intrinsically selected by the machine learning-driven models. The application is available either as a standalone version through GitHub or as webserver (www.edtox.fi). A comprehensive tutorial on running the application is available as Supplementary File S1. The case study with EDTox R Shiny application is fully documented in Supplementary File S2. The Supplementary File S3 provides a summarized table of the type of input and output required in each module of the application.

Acknowledgements

The authors wish to acknowledge CSC—IT Center for Science, Finland, for computational resources.

Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant 825762.

Conflict of Interest: Dr. Amirhossein Sakhteman implemented the first version of the proposed software when working at University of Eastern Finland as project researcher for the project EDCMET.

Data Availability

The data underlying this article are available in Zenodo at https://doi.org/10.5281/zenodo.5817093, and can be accessed with 10.5281/zenodo.5817093.

References

Davis,A.P. *et al.* (2019) The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.*, **47**, D948–D954.

Davis,A.P. *et al.* (2021) Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, **49**, D1138–D1143.

Filer,D. *et al.* (2014) Test driving ToxCast: endocrine profiling for 1858 chemicals included in phase II. *Curr. Opin. Pharmacol.*, **19**, 145–152.

Ganter,B. *et al.* (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.*, **119**, 219–244.

Igarashi,Y. *et al.* (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.*, **43**, D921–D927.

Sakhteman,A. *et al.* (2021) A toxicogenomic data space for system-level understanding and prediction of EDC-induced toxicity. *Environ. Int.*, **156**, 106751.

Subramanian,A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.e17.