OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Case Study

A systems biology approach for grouping compounds [**SysGroup**]

SUMMARY

This case study will use the approach of the diXa / DECO2 (Cefic-LRI AIMT4[1]) projects to reproduce and extend the results obtained on the identification of hepatotoxicant groups based on similarity in mechanisms of action (omics-based) and chemical structure using services from OpenRiskNet. Figure 1 displays the workflow that was used in this case study.

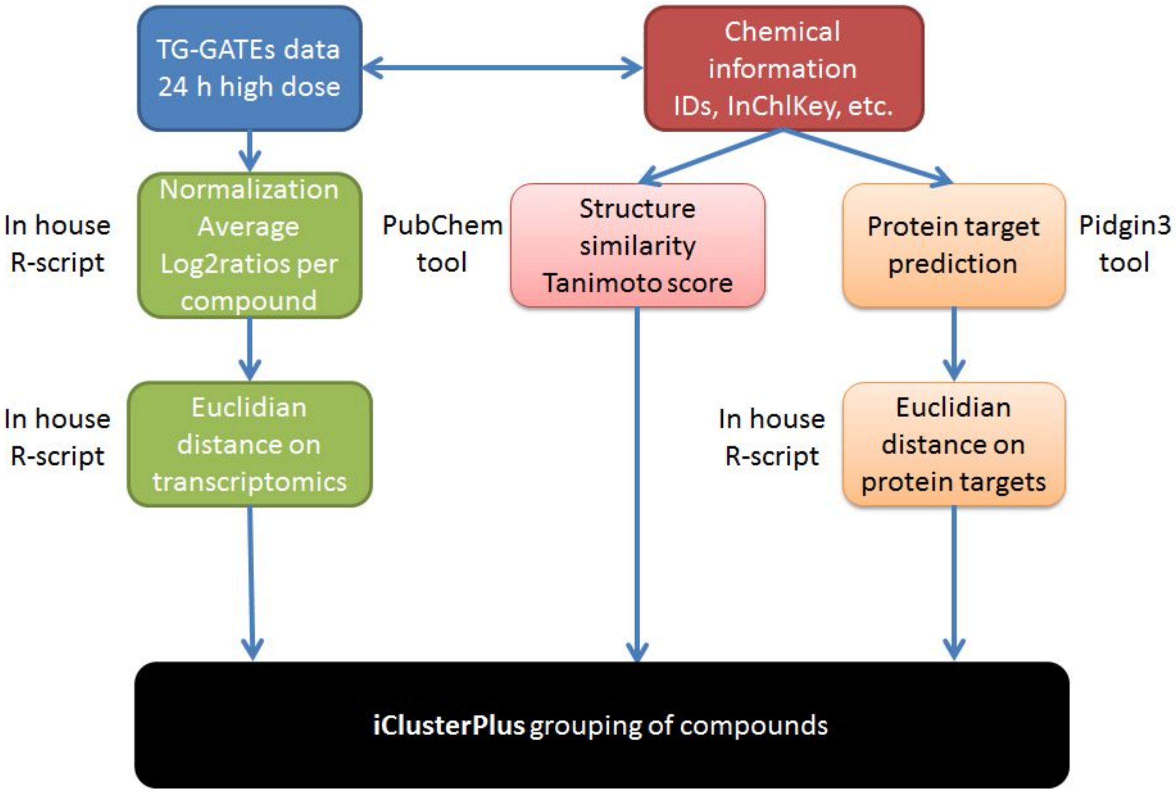

Figure 1. Workflow for grouping compounds by integrating transcriptomics data, Tanimoto similarity scores and ligand-based target predictions using iClusterPlus.

[1] http://cefic-lri.org/projects/aimt4-um-deco2-moving-from-deco-towards-oecd/

DESCRIPTION

Implementation team

Coordination:

- Danyel Jennen, Maastricht University, Department of Toxicogenomics

Other members:

- Jumamurat Bayjanov, Maastricht University, Department of Toxicogenomics

Case Study objective

The objective of this case study is to implement an integrated analysis using chemoinformatics and omics data for improved grouping of compounds with similar toxicity and/or mode of action.

Risk assessment framework

SysGroup covers the identification of use scenario / chemical of concern / collection of existing information (Tier 0 in the selected framework) and its steps related to:

- Identification of molecular structure;
- Collection of support data;
- Identification of analogues / suitability assessment and existing data.

DEVELOPMENT

Databases and tools

In this case study TG-GATEs transcriptomics data from primary human hepatocytes exposed to 139 compounds for 24h and the highest dosage were used. This dataset was obtained from the OpenRiskNet service “Transcriptomics data from human, mouse, rat in vitro liver models”[2].

For the 139 compounds PubChem[3] was used to obtain 2D Tanimoto scores and PIDGIN3[4] was used to retrieve ligand-based target predictions.

Integration of the transcriptomics data with the chemoinformatics data was performed using iClusterPlus[5], an integrative clustering framework developed to integrate diverse data types (i.e. binary, categorical, and continuous values) by a latent variable approach.

Technical implementation

Integration with other case studies is needed. SysGroup acquires information and data from the DataCure or TGX case study and can feed into AOPLink and ModelRX.

Through the services ToxPlanet and ToxicoDB of the Implementation Challenge winners Toxplanet and UHH, respectively, information on the obtained groups of chemicals is obtained.

Currently available services:

- Transcriptomics data from human, mouse, rat in vitro liver models
    - Repository for transcriptomics data from multiple in vitro human, rat and mouse toxicogenomics projects
    - Service type: Database / data source

[2] https://openrisknet.org/e-infrastructure/services/164/
[3] https://pubchem.ncbi.nlm.nih.gov/score_matrix/score_matrix.cgi
[4] https://pidginv3.readthedocs.io/en/latest/
[5] https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html

OUTCOMES

The outcome of this case study is the workflow as shown in Figure 1. This workflow was setup using GNU Make[6] and is available through OpenRiskNet's GitHub[7].

The workflow contains several steps to integrate 3 different datasets or data types for the purpose of grouping chemicals based on similarity in mechanisms of action (omics-based) and their chemoinformatics properties.

Step 1: Obtaining transcriptomics data

- Transcriptomics data were obtained from the OpenRiskNet service "Transcriptomics data from human, mouse, rat in vitro liver models"[8]. Here we only selected normalized data of from TG-GATEs (i.e. from primary human hepatocytes exposed to 139 compounds for 24h and the highest dosage);
- Scaled Euclidean distances (scale 0-1; 0 = most similar;1 = most dissimilar) were calculated from these transcriptomics profiles;

→ Result step 1: a 139 x 139 distance matrix for transcriptomics data with ChEMBL ID as identifier.

Step 2: Calculating a Tanimoto score for each compound

- Convert ChEMBL ID to InChIKey using ChEMBL conversion tool and subsequently InChIKey to CID using PubChem conversion tool;
- Use the list of CID's to calculate 2D Tanimoto scores via PubChem tool, which are shown as percentages;
- Convert percentages to 0-1 scale;
- Convert CID to ChEMBL ID;
- Convert 2D Tanimoto scores to a distance, by subtracting Tanimoto score from 1;

→ Result step 2: a 139 x 139 distance matrix for 2D Tanimoto scores with ChEMBL ID as identifier.

Step 3: Predicting ligand-based protein targets for each compound

- Convert ChEMBL ID to smiles ChEMBL conversion tool;
- Predict ligand-based protein targets from smiles using Pidgin3 tool;
- Scaled Euclidean distances (scale 0-1; 0 = most similar;1 = most dissimilar) were calculated from these protein target profiles;
- Convert smiles to ChEMBL ID;

→ Result step 3: a 139 x 139 distance matrix for ligand-based protein targets with ChEMBL ID as identifier.

Step 4: Grouping of chemicals from integrated data

- For the obtained matrices from step 1-3 the order of the ChEMBL ID's are set in the same order;
- The 3 matrices are integrated using iClusterPlus; the number of clusters is set to 46 (~1/3 of 139);

→ Result step 4: heatmaps for each data type sorted per cluster (see Figure 2).

[6] https://www.gnu.org/software/make/
[7] https://github.com/OpenRiskNet/notebooks/tree/master/openrisknet_sysgroup
[8] https://openrisknet.org/e-infrastructure/services/164/

The grouping of the chemicals in these clusters needs to be further investigated using toxicological data from the DataCure case study and various OpenRiskNet services, such as **ToxicoDB** and **ToxPlanet**. However, due to time constraints this final step was not achieved within the time frame of the project.

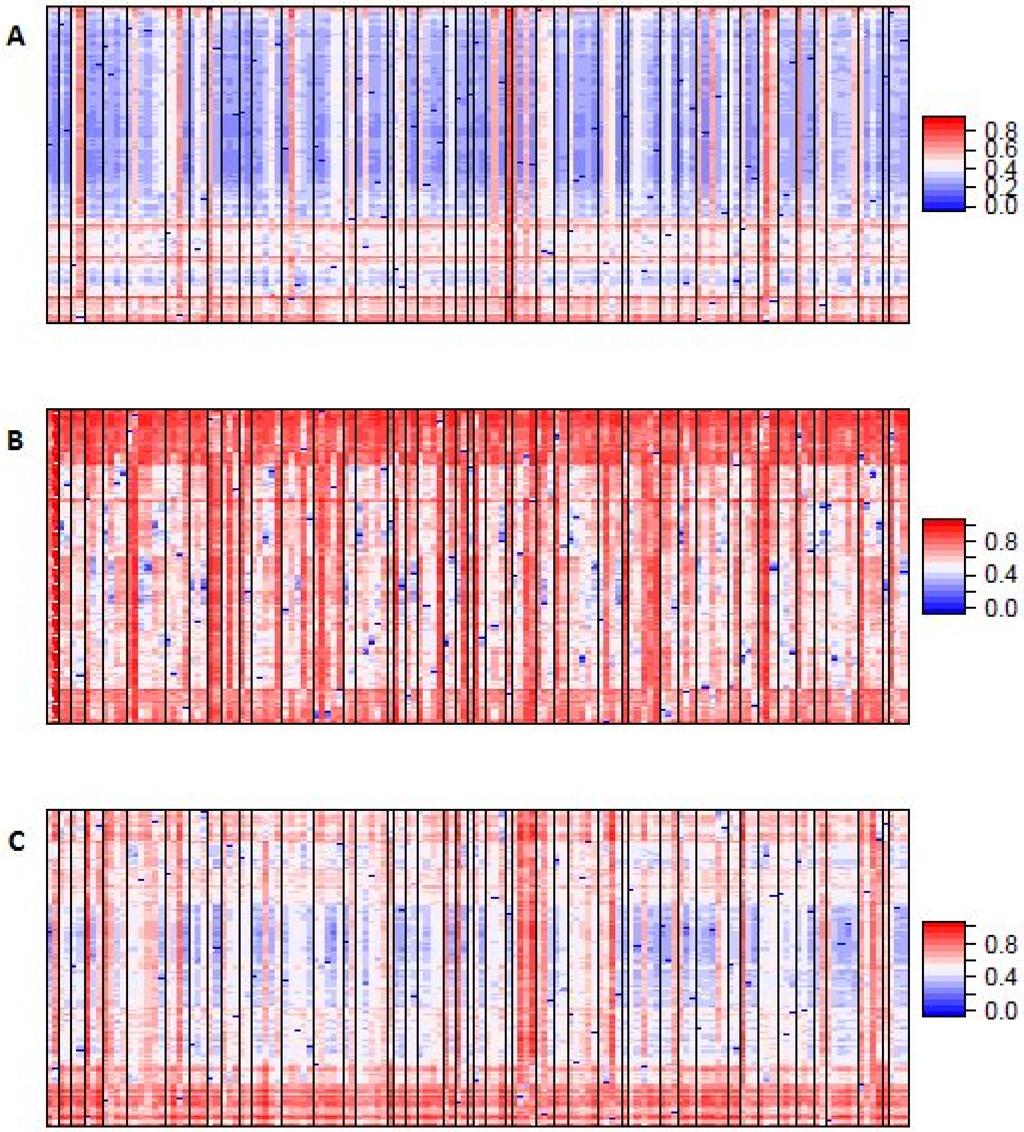

Figure 2. iClusterPlus result showing the grouping of 139 compounds in 46 clusters; A) transcriptomics data, B) 2D Tanimoto scores, and C) ligand-based protein predictions

Related resources

OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis

Marvin Martens, Thomas Exner, Nofisat Oki, Danyel Jennen, Jumamurat Bayjanov, Chris Evelo, Tim Dudgeon, Egon Willighagen

28 August 2019 | Poster