OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

TGX

Toxicogenomics-based prediction and mechanism identification

SysGroup

A systems biology approach for grouping compounds

Danyel Jennen

OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and *in silico* Analysis and Modelling in Risk Assessment
Project Number 731075

Toxicogenomics based prediction and mechanism identification [TGX]

CS leader: Danyel Jennen (UM), **Involved:** UM, VU, CRG

Risk Assessment Framework
Tier 0.3-0.4 (data collection), 1.6 (MOA)

AIM: To provide a transcriptomics-based hazard prediction model for identification of specific molecular initiating events (MIE)

The foreseen transcriptomics-based hazard prediction model will be applied based on:

(A) top-down Creation of prediction models based on differentially regulated genes

(B) bottom-up Using knowledge of stress response pathways to integrate data sets for their activation or inhibition (bottom-up approach).

The MIEs can include, but are not limited to:

(1) Genotoxicity (p53 activation), (2) Oxidative stress (Nrf2 activation), (3) Endoplasmic Reticulum Stress (unfolded protein response), (4) Dioxin-like activity (AhR receptor activation), (5) HIF1 alpha activation and (6) Nuclear receptor activation (e.g. for endocrine disruption).

Databases

- diXa / BioStudies (UM)
- TG-GATEs
- EU-ToxRisk (nascent)
- HeCaToS (nascent)
- ArrayExpress / GEO

Tools / APIs

- top-down: Data normalisation tools, prediction tools such as Caret;
- bottom-up: ToxPi

Service integration

- Service integration will be needed for the omics databases; knowledge bases and data mining; processing and analysis.

Activities

- First top-down case study based on Magkoufopoulou *et al* 2012 paper
- Second top-down case study on meta-analysis for genotoxicity prediction in human, rat and mouse in vitro cell models

https://openrisknet.org/e-infrastructure/development/case-studies/case-study-tgx/

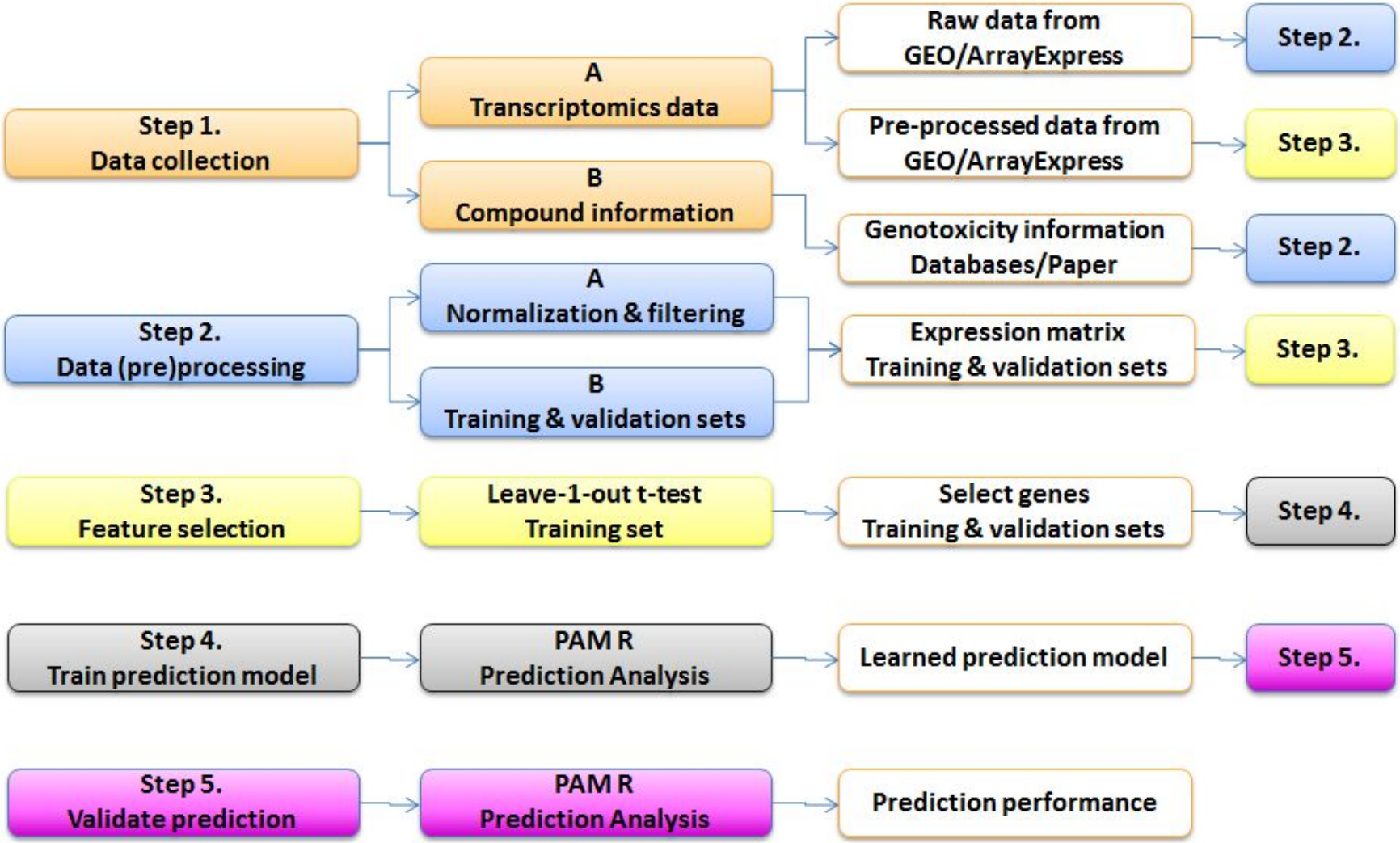
Top-down approach → Case study 1

Create workflow based on Magkoufopoulou *et al.* 2012

Carcinogenesis vol.33 no.7 pp.1421–1429, 2012
doi:10.1093/carcin/bgs182
Advance Access Publication May 23, 2012

A transcriptomics-based *in vitro* assay for predicting chemical genotoxicity *in vivo*

C.Magkoufopoulou[1,2], S.M.H.Claessen[1], M.Tsamou[1], D.G. J.Jennen[1,2], J.C.S.Kleinjans[1,2], J.H.M.van Delft[1,2,*]

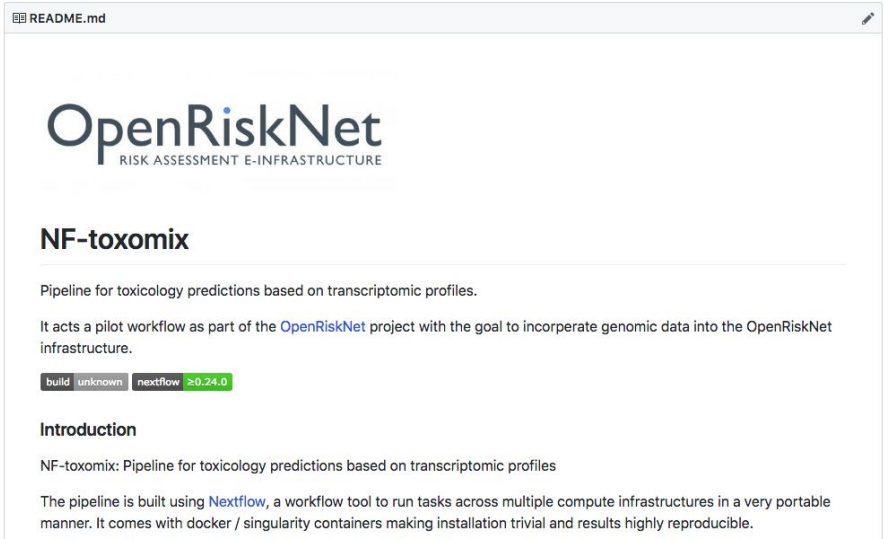

Step 1.
Data collection
A
Transcriptomics data
B
Compound information
Raw data from
GEO/ArrayExpress
Pre-processed data from
GEO/ArrayExpress
Genotoxicity information
Databases/Paper
Step 2.
Step 3.
Step 2.
Step 2.
Data (pre)processing
A
Normalization & filtering
B
Training & validation sets
Expression matrix
Training & validation sets
Step 3.
Step 3.
Feature selection
Leave-1-out t-test
Training set
Select genes
Training & validation sets
Step 4.
Step 4.
Train prediction model
PAM R
Prediction Analysis
Learned prediction model
Step 5.
Step 5.
Validate prediction
PAM R
Prediction Analysis
Prediction performance

Current status

Workflow has been established in Snakemake and is available via Gitlab

Workflow was converted into NextFlow pipeline;

Converted into a generic workflow applicable to other datasets.


README.md

OpenRiskNet
RISK ASSESSMENT E-INFRASTRUCTURE

NF-toxomix

Pipeline for toxicology predictions based on transcriptomic profiles.

It acts a pilot workflow as part of the OpenRiskNet project with the goal to incorperate genomic data into the OpenRiskNet infrastructure.

build unknown nextflow ≥0.24.0

Introduction

NF-toxomix: Pipeline for toxicology predictions based on transcriptomic profiles

The pipeline is built using Nextflow, a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It comes with docker / singularity containers making installation trivial and results highly reproducible.

Generic workflow

- Data collection from GEO/ArrayExpress → API available
- Retrieve (geno)toxicity information
  - from databases
    - possible, but depends of required data
    - expert knowledge needed
  - from paper
    - difficult as presentation of information may vary
    - online, pdf → difficult, because of format;
      suppl. data as Word, Excel, txt files → possible
- All other steps comprise of R-scripts → easy to adapt

→ data collection relies on available metadata, used format & ontologies

Collaboration with implementation challenge winners

toxplanet

Find the chemical hazard and toxicology literature you need with our information portal

Our Federated Search engine searches the content from 500+ websites and quickly delivers relevant literature.

UHN

TOXICODB

Investigate the pathways triggered by exposure to toxic substances (coming soon)

PDF: Table detection and -segmentation

Fraunhofer SCAI

Top-down approach → Case study 2

Meta-analysis for genotoxicity prediction in human, rat and mouse *in vitro* cell models

In this case study transcriptomics data obtained from multiple data sources will be used to build a prediction model for *in vivo* genotoxicity.

A similar approach will be applied as in case study 1.

Preliminary results on the rat data have been presented at the ICCA-LRI workshop 2018, Ottawa, Canada
Preliminary results on the human data have been presented at EUROTOX 2018, Brussels, Belgium

Current status

The generic workflow will be applicable in this case study as well.

- Additional databases have been used, e.g. the diXa Data Warehouse
  → included in EBI's BioStudies → an API is available
- In vitro liver transcriptomics data from multiple human, rat & mouse cell models is available via Gitlab
- (Geno)toxicity data on all compounds has been collected
  → valuable source for ToxicoDB

Other services from the implementation challenge can be incorporated.
→ ToxPlanet databases potentially can provide genotoxicity information

A systems biology approach for grouping compounds [SysGroup]

CS leader: Danyel Jennen (UM), **Involved:** UM, Fraunhofer, CRG

AIM: To provide the services for **improved grouping** of compounds by integrating chemoinformatics and omics data.

Will use the approach of the diXa / DECO2 (Cefic-LRI AIMT4) 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.

Objectives

- This case study will implement an integrated analysis approach using chemoinformatics and omics data for improved grouping of compounds with similar toxicity and/or mode of action

Risk Assessment Framework
Tier 0.2-0.4 (data collection, read across)

Databases

- diXa /BioStudies (UM)
- PubChem
- ChEMBL

Tools / APIs

- PubChem
- ChEMBL or PIDGIN
- (pre)processing tools for gene expression data (e.g. microarray data) → e.g. arrayanalysis.org (UM)
- iClusterPlus

Required steps

- Chemical similarity calculated by 2D or 3D Tanimoto coefficient (PubChem)
- Protein target prediction (ChEMBL/PIDGIN)
- Interface to diXa for obtaining gene expression data
- Integration of the multiple data sources and grouping by iClusterPlus

Service integration

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

Flowchart SysGroup