HARNESSING BIG DATA FOR CHEMICAL TOXICITY PROFILING

Rohit Bhatia • People for the Ethical Treatment of Animals, PO Box 28260, Juhu, Mumbai, India • RohitB@petaindia.org

INTRODUCTION

- Traditional approaches to assessing the toxicity of chemicals involve expensive and time-consuming tests on animals, the results of which have limited applicability to humans.
- In recent years, scientists have developed and validated many animal-free methods that can predict human health and environmental effects. These approaches are often based on human cells and tissues and may include high-throughput screening (HTS) methods and advanced *in silico* models.
- Several international data-sharing projects have resulted in large amounts of publicly available toxicity data that can now be used to predict the toxicity of chemicals without conducting tests on animals.
- The enormous size of these databases makes them difficult to process using traditional data-analysis tools such as SQL, SAS, R, and Excel. However, recent advances in data science and big-data analytics offer new methods for data-driven predictions of chemical toxicity.

BIG DATA AND TOXICITY RESEARCH

- "Big data" is a term that describes large volumes of data – including both structured and unstructured. The "Four V's" – volume, velocity, variety, and veracity – are used to determine whether or not the data set can be considered big data.
- The data-information-knowledge-wisdom (DIKW) pyramid can be used for making decisions based on big data.
- Publicly available data repositories such as PubChem, ChEMBL, TOXNET, Seurat, REACH, and CEBS provide a huge amount of data which can be used to predict the toxicity of a substance. (See Table 1.)

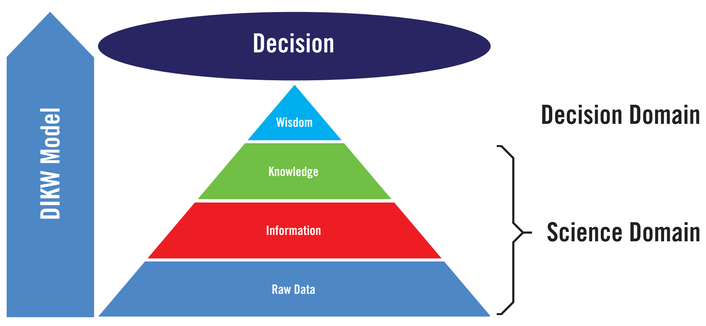

Figure 1: DIKW pyramid: from big data to decision-making[1]

Database	Description	Type of Data
PubChem	89 million compounds 23 billion data points	Toxicity, pharmaceutical, genomic, and literature
ChEMBL	6 million compounds 3.3 million data points	Literature
TOXNET	50,000 environmental compounds	*In vitro* and *in vivo*
CEBS	10,000 toxicity bioassays	Gene expression

Table 1: Data available via PubChem, ChEMBL, TOXNET, and CEBS[2]

USING PUBCHEM DATA FOR PREDICTING TOXICITY

- PubChem is a publicly available chemical information archive developed by the US National Institutes of Health that organises data into three primary databases: Substance, Compound, and BioAssay.
- It contains 219 million substances, 89 million compounds, and the results from 1 million assays. It can automate for virtual screening using the following tools: Entrez Utilities (also called E-utilities or E-utils), Power User Gateway (PUG), PUG SOAP, and PUG REST.
- Zhu *et al* predicted the acute animal toxicity of compounds using their bioactivity profiles extracted from data in the PubChem BioAssay database.[3]
- Kim *et al* developed a model to predict oxidative stress-induced hepatotoxicity using HTS data archived in PubChem.[4]

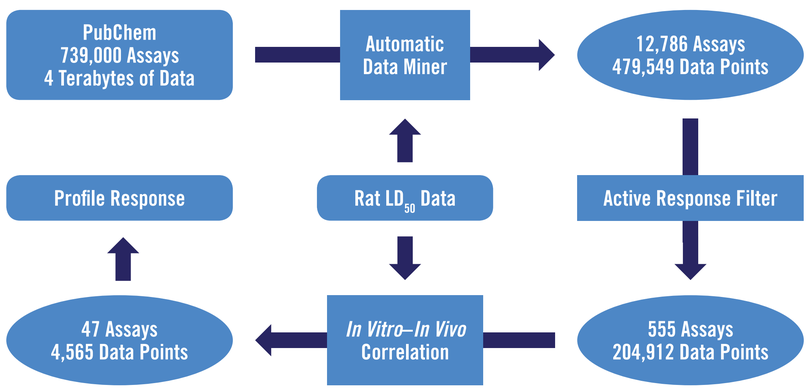

Figure 2: Model workflow for predicting animal acute toxicity using LD_{50} values of rats[5]

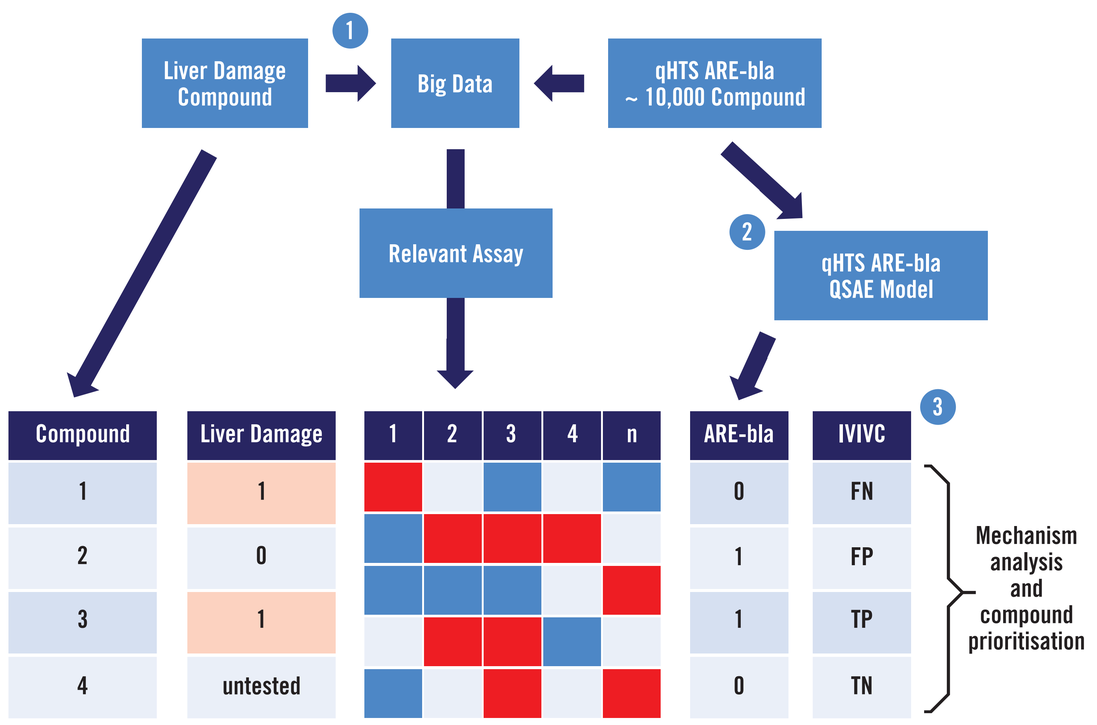

Figure 3: The three-stage workflow model for predicting liver toxicity is as follows: Stage 1: automated biological response profiling. Stage 2: quantitative structure-activity relationship modelling of the HTS antioxidant response element beta-lactamase reporter gene. Stage 3: evaluation of chemical *in vitro–in vivo* correlation. In the columns, active is red and "1"; inactive is blue and "0"; and inconclusive or untested is blank.[6]

RECOMMENDATIONS

- Researchers should become familiar with and use open-access data to develop models for predicting the toxicity of chemical compounds.
- Regulatory agencies such as the Central Insecticides Board & Registration Committee and the Central Drugs Standard Control Organization should share available data with researchers to facilitate the development and validation of predictive toxicological tools.
- The government should establish a national computational facility for the validation of models for predicting the toxicity of chemical compounds.

REFERENCES

[1]Lokers R, Knapen R, Janssen S, et al. Analysis of big data technologies for use in agro-environmental science. *Enviro Mod & Soft*. 2016;84:494-504.
[2]Zhu H, Zhang J, Kim MT, et al. Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem Res Toxicol*. 2014;27(10):1643-1651.
[3]*Ibid*.
[4]Kim MT, Huang R, Sedykh A, et al. Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ Health Perspect*. 2016;124:634-641.
[5]Zhang J, Hsieh JH, Zhu H. Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology. *PLoS ONE*. 2014;9(6):11-15.
[6]Sedykh A, Zhu H, Tang H, et al. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ Health Perspect*. 2011;119(3):364-370.