Integration into Big Data: First Steps to Support Reuse of Comprehensive Toxicity Model Modules

Sean Watford[1,5], Jeff Edwards[2], Ivan Rusyn[3], David Reif[4], Richard Judson[5], Matt Martin[5]

1 UNC-Chapel Hill, 2 ICF International, 3 Texas A&M, 4 NCSU, 5 USEPA/ORD/NCCT

Sean Watford | watford.sean@epa.gov | 919-541-7655

United States Environmental Protection Agency

Abstract

Data surrounding the needs of human disease and toxicity modeling are largely siloed limiting the ability to extend and reuse modules across knowledge domains. Using an infrastructure that supports integration across knowledge domains (animal toxicology, high-throughput screening, genomics, proteomics, disease, exposure, product use, chemistry, etc.) increases the ability to evaluate, extend and expand models. For example, type II diabetes is a metabolic disorder caused and influenced by a combination of genetics, lifestyle and environment. In order to quantify the contribution of each factor and related confounders (e.g., diagnosis, screening, and treatment), the modeling framework relies on the ability to systematically access information across many knowledge domains to more accurately resolve the uncertainty resulting from the complexity within and across each factor. A first step to developing an integrated system was to develop an object model (i.e., a conceptual representation of each knowledge domain; ontologies) to resolve data redundancy and granularity issues from the complexity of the data. The advantage of an object model over siloed databases was the ability to confidently link and merge previously disconnected datasets. The current object model enables the modular development of systems capable of providing an extensible framework for building a more comprehensive human disease model.

Objectives

- Develop an integrated network of toxicity information to foster data exploration and hypothesis generation
- Steps
  - Identify data resources covering relevant domains of knowledge
  - Identify biomedical ontologies that can be used as a standard for each data source
  - Map data to ontologies building an integrated network

Domains of Knowledge

Figure 1: Six Overlapping Domains of Knowledge

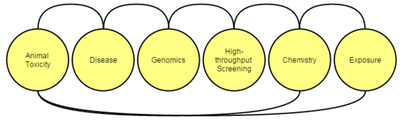

Figure 1: Six major domains of knowledge relevant to toxicity. Each domain of knowledge contains overlapping concepts with another domain. For example, exposure and high-throughput screening both have overlapping concepts with chemistry. Combing concepts from each domain will provide more comprehensive insight into toxicity. Full coverage of toxicity concepts are not exclusively limited to the above six.

Data Sources

Table 1: Publicly Available Data Sources Covering Domains of Knowledge

Source	Domains
ToxCast: Toxicity Forecaster	High-Throughput Screening; Chemistry; Genomics
ToxRef: Toxicity Reference Database	Animal Toxicity; Chemistry
PubChem	Chemistry; High-Throughput Screening; Animal Toxicity
CPCat: Chemical and Product Categories	Chemistry; Exposure
ExpoCast: Exposure Forecasting	Chemistry; Exposure
DSSTox: Distributed Structure-Searchable Toxicity	Chemistry
CTD: Comparative Toxicogenomics Database	Chemistry; Genomics

Table 1: Relevant publicly available resources for comprehensive coverage of biological interactions that cover domains of knowledge from Figure 1.

Ontologies

Figure 2: Biomedical Ontologies Connected Through Mappings

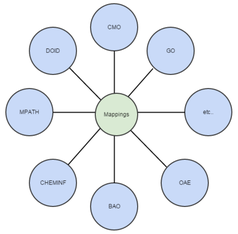

Figure 2: Ontologies comprise the basis of data integration through standardization of concepts across multiple domains of knowledge. Concepts are shared across ontologies through mappings creating a large unstructured network to define an integrated data space. Concepts defined within an ontology can either be fully adopted or extended to specifically fit a particular domain.

Table 2: Names and Descriptions of Biomedical Ontologies Relevant to Toxicity Data Sources

Ontology	Description
CMO: Clinical Measurement Ontology	Standardizes morphological and physiological measurement records from model organisms
GO: Gene Ontology	Represents biological process, cell functions, and cell components related to genes
BAO: Bioassay Ontology	Represents chemical biology screening assays and the results
MPATH: Mouse Pathology Ontology	Represents mouse pathology phenotypes
CHEMINF: Chemical Information Ontology	Represents a collection of cheminformatics descriptors
OAE: Ontology of Adverse Events	Standardizes reporting of adverse events
DOID: Human Disease Ontology	Represents human disease with a hierarchical controlled vocabulary
ExO: Exposure Ontology	Represents environmental exposure concepts

Table 2: These ontologies standardize the representation of concepts from each of the data domains from Figure 1. Data from each of the data sources from Table 1 can be mapped to the above ontologies for integration (illustrated by Figure 3).

Data Integration

Figure 3: Integrated Network Toxicity Resources Connected via Ontology Mappings

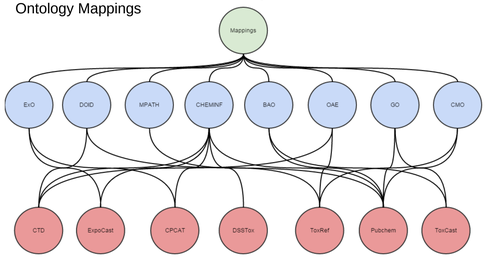

Figure 3: Data from each data source in Table 1 can be mapped to one or more ontologies from Table 2. Due to the connected concepts via ontology mappings, an integrated network of toxicity concepts backed with data is created and available to browse, analyze, and investigate.

Type II Diabetes Example

Figure 4: Individual Associations Discovered from Manual Search

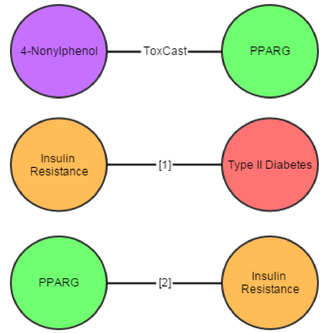

Figure 4: Shown are direct associations between the chemical 4-Nonylphenol, PPARG (peroxisome proliferator-activated receptor gamma) gene, insulin resistance, and type II diabetes found by performing a cursory, manual search through literature and ToxCast data. ToxCast shows 4-Nonylphenol as a hit across a PPARG assay. Insulin resistance is associated with type II diabetes as a disease phenotype. PPARG has been shown to play a role in insulin sensitization.

Figure 5: Direct and Inferred Associations

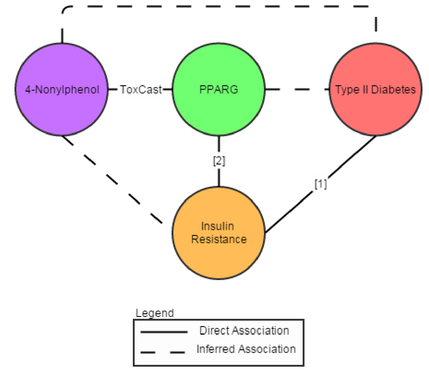

Figure 5: Associations between 4-Nonylphenol and type II diabetes can be inferred from the direct associations between seemingly independent entities surrounding the disease. The strength of the associations can be measured, and, although a direct association linking chemical exposure to a disease, may be difficult to obtain, a hypothesis about possible contributions or even the mechanistic basis can be generated.

References

1. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2010;33(Suppl 1):S62-S69. doi:10.2337/dc10-S062.
2. Kintscher, U, Law, RE. PPARgamme-mediated insulin sensitization: the importance of fat versus muscle. American Journal of Physiology Endocrinology and Metabolism. 2005;288(2):E287-91.

Summary and Future Work

- Mapping data sources to ontologies covering concepts relevant to toxicity will create an integrated network of publicly available data for browsing, analyses, and investigation
- Next Steps
  - Create models to analyze the overlap and full coverage of concepts relevant to toxicity.
  - Continue to expand the network for larger coverage of largely overlapping biological and biomedical fields.
  - Design and implement the preceding concepts for public consumption (see architecture concept in Figure 6).

Architecture Concept

Figure 6: Overview of Architecture Concept for Public Access

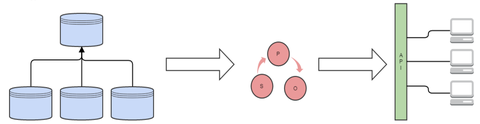

Figure 6: ETL (extract, translate, and load) operations maintain normalization, versioning, and prevent redundancy. ETL operations convert data into triples. API (Application Programming Interface) is created for data access. Clients include web applications, analytic applications, etc.

Innovative Research for a Sustainable Future