

From genes, proteins, metabolites, tissue analytics and literature towards a coherent knowledgebase: Semantics at work

Erich A. Gombocz¹⁾, Robert A. Stanley¹⁾, Toshiro Nishimura¹⁾, Chuck Rockey¹⁾

¹IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, U.S.A. [Correspondence: egombocz@io-informatics.com]

Summary

To gain understanding of complex pharmacodynamic responses to a treatment in respect to its beneficial and adverse effects within the organism has been extremely difficult. Experimental observations alone – even when obtained from different experimental techniques – cannot be meaningfully related to each other, let alone to systems biological functions.

In this work, we use a data integration approach which aims to remedy such obstacles. This is done through coherent merging of findings from gene expression profiling, proteomics, metabolite analysis and tissues analytics into a network representation in which data have been classified and categorized in regards to their relationships to each other. The common ontology knowledgebase then is further enhanced through integration with public domain knowledge from reference resources (NCBI Entrez' set of databases, UniProt, IntAct, BioGrid, KEGG, HMDB) and mined literature to assist in mapping of multi-pathway relationships to experimental observations. To make such a technique generally applicable, requires recognition and reconciliation of different existing taxonomies, incompatible semantic data models and standards and to consolidate vocabularies for data classes, terms and their relationships. A "smart" data taxonomy merge also needs to account for hierarchical adjustments into super- or subclasses of the new, merged ontology tree whenever needed. The resulting ontological, graphical visualization can then be used to describe, link, browse and search potential biomarkers according to their correlative responses to perturbation, within a systems biology environment.

Genetic and metabolic changes typically occur at lower doses and earlier times than those that produce pathological changes. Strength and broader applicability of semantics to link pharmacodynamic correlations functionally within the biochemical network are discussed using examples of multi-OMICS toxicity responses in different tissues and across several known chemotoxicants. The ability to deduce knowledge about biological processes and their intertwined interactions and its applicability even in cases of different experimental models are demonstrated.

Challenges

- Metabolic and gene expression changes may result from same toxic insult despite representing very different biological processes.
- There is no direct relationship between pharmacodynamic correlations and functions in biological systems networks.
- General data coherence: source data categorized in different taxonomies; normalization issues, term inconsistencies and property disparities across large, complex experimental data sets
- In many cases, data relationships are not a priori contained in the data sets
- Complexity and lack of intuitive, science-driven tools for network analysis makes such approaches non-appealing to researchers.
- Multiple, incompatible semantic standards make merging towards a common ontology extremely difficult; this also impacts applicability of query and reasoning.
- Scalability and performance issues confine most network approaches to relatively small datasets.

Experimental Models

Hepatotoxicity study

- Panel of several hepatotoxicants, single oral dose (placebo, low, mid, high) in groups of 4 rats, at 6, 24 and 48 hrs.
- Metabolomic analysis of liver, serum and urine (1603 metabolites).
- Microarray analysis of liver and whole blood (31096 transcript probes).

Alcohol study

- High doses t.i.d. for four days, with and without 24h withdrawal
- Metabolic analysis of plasma, liver and brain
- Microarray analysis of liver and brain

Process

- Identify metabolites and genes with significant perturbation (LC-MS and GEP analysis).
- Select robust correlations between independent analytical results
- Map results into a semantic framework to visualize, investigate and analyze data relationships.
- Associate significant elements of those networks with reference data sources
- Use thesauri to consolidate data class and relationship synonyms, and combine experimental data with literature
- Map pathway enzymes from public sources to experimental data and merge into a common ontology network across taxonomies and standards
- Apply criteria such as weighing and connection depth to reduce network complexity for visualization.
- Perform graphical, textual and SPARQL queries to specify multi-parametric conditions (such as time/dose dependencies), and re-plot the results as sub-networks to qualify potential biomarker panels.
- Output the knowledgebase to RDF, N3 or triples store backend for inference, reasoning and iterative improvements of the underlying model.

Applicability

- Decode underlying mechanisms of complex biological functions through Visualization, exploration and network analysis of experimental observations in context.
- Optimize therapeutic effect and minimize toxicity in disease treatment.
- Generate knowledge about biological processes from different experimental models to validate the best animal model for complex diseases.

Results

- Merging of multi-modal datasets into a common ontology knowledgebase and association with canonical reference sources provides insights in complex processes on the organism level.
- The method of semantic integration can be applied across different experimental models.
- The presented network viewer can comfortably handle one million assertions in memory (at 2GB of RAM), and is able to directly query the knowledgebase.

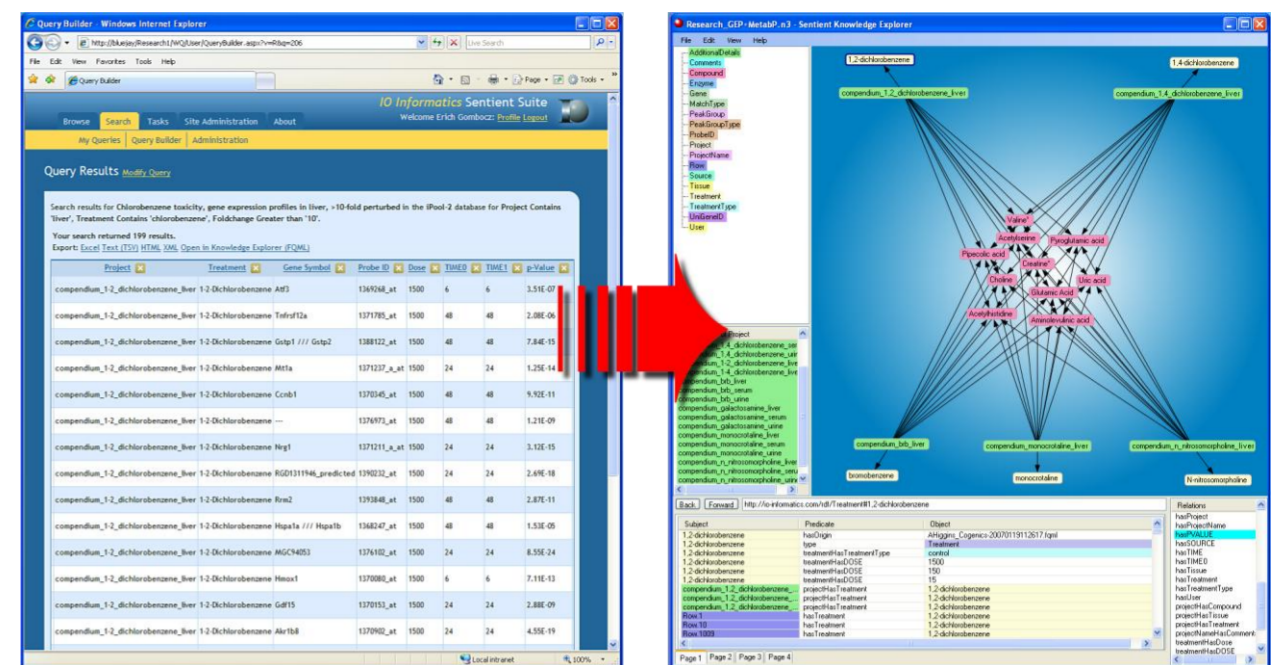


Fig.1: Semantic data merging: Query gene expression (left), merging with metabolic changes into a common ontology in Sentient Knowledge Explorer™. Common metabolic changes across toxicants are easily revealed (right).

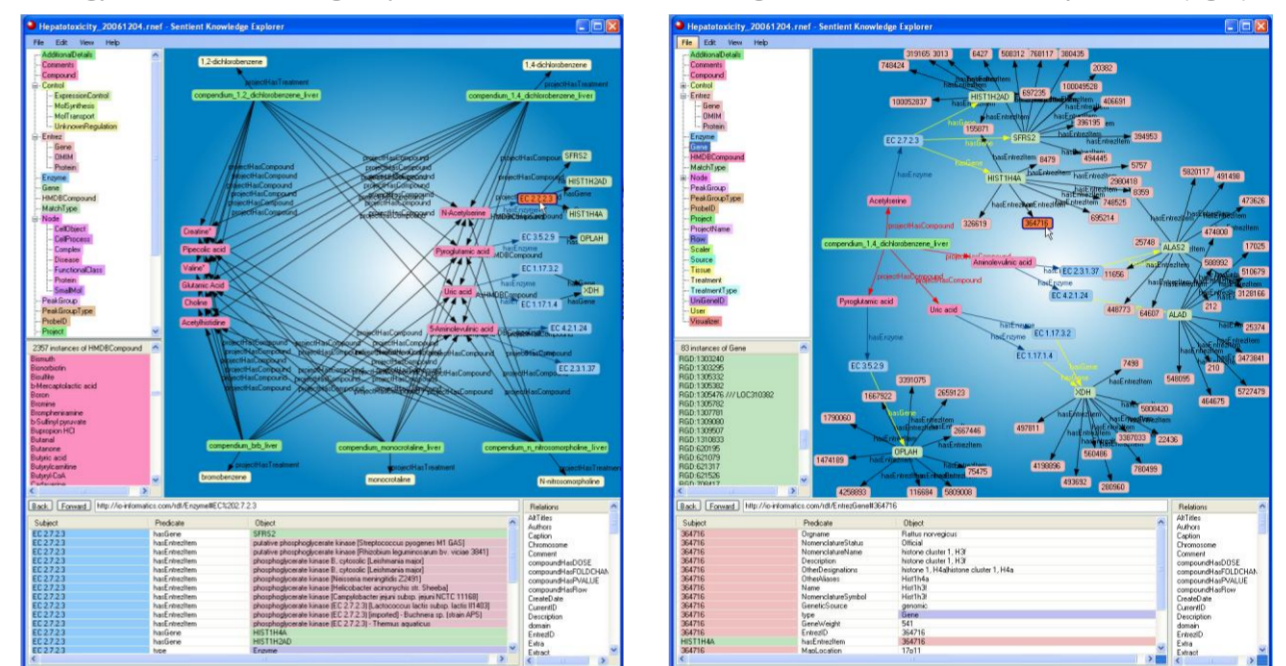


Fig.2: Integrating reference resources: Drill out to HMDB and NCBI's Gene, Protein and OMIM and mined literature for cellular processes and pathways (left). Entrez Gene is mapped to experimental gene expression profiling (right).

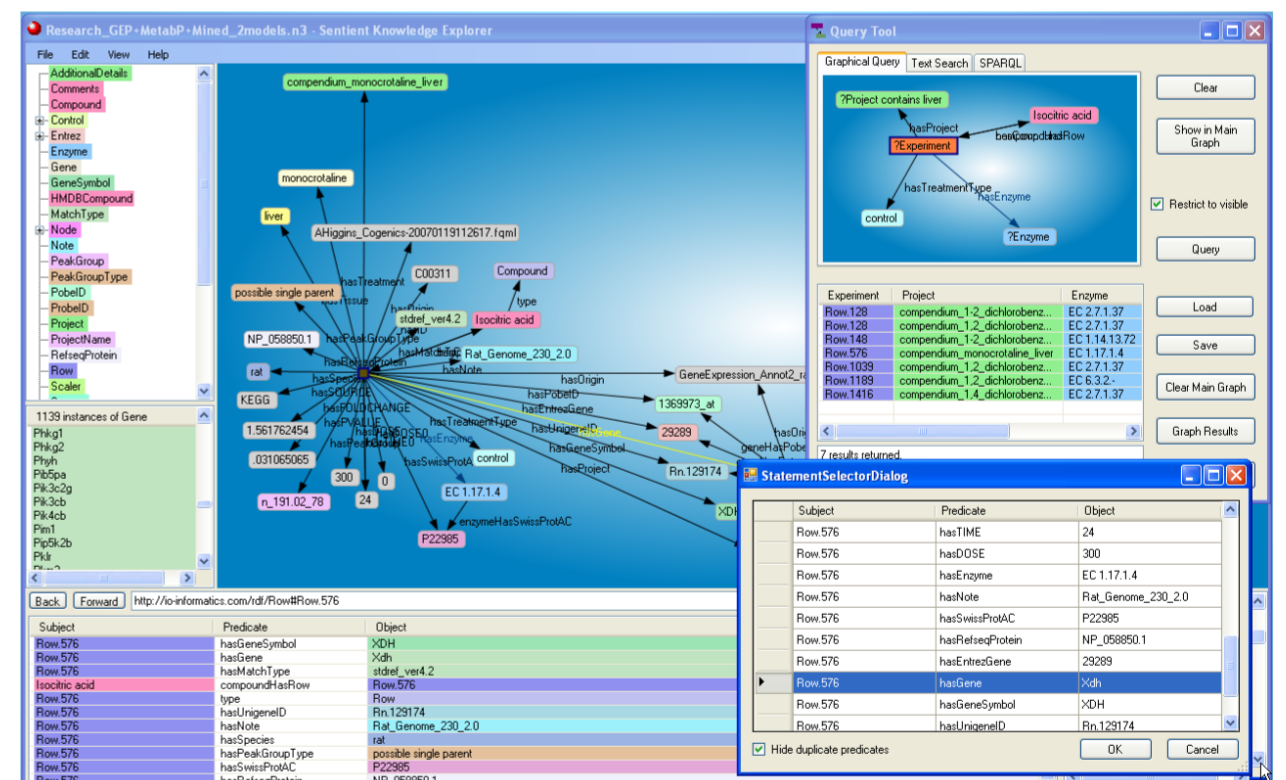


Fig.3: Biological mechanisms revealed across experimental models: Complexity reduction through graphical queries.

Conclusions

Using a semantic, common ontology framework to explore toxic perturbations in large metabolic and genomic datasets across several toxicants and different tissues, we were able to:

- Correlate metabolites and genes across different treatments to ascertain commonality of effects for a class of drugs.
- Review effects across tissues to find common metabolic markers in the most accessible tissue for diagnostics.
- Validate changes in biomarkers associated with known common mechanisms of toxicity (oxidative stress [Glutathion metabolism], liver function [Bile acid and Urea cycle], Ketoacidosis).
- Detect biological similarities even in cases of different experimental models.
- Discover new pharmacodynamically and biologically linked components and investigate their functional relevancy.

Acknowledgements

This work was conducted under NIST Advanced Technology Program (ATP), Award # 70NANB2H3009 as a Joint Venture between Icoria / Cogenics (Division of CLDA) and IO Informatics. Microarray studies were conducted under NIEHS contract # N01-ES-65406. The Alcohol study was conducted under NIAAA contract # HSN281200510008C. This work was also made possible through insights from members of IO Informatics' Working Group on "Semantic Applications for Translational Research". The authors would like to acknowledge Ted Slater (Pfizer), Jonas Almeida (MD Anderson Cancer Center), Pat Hurban (CLDA), Alan Higgins (Viamet Pharmaceuticals) and Bruce McManus and Mark Wilkinson (both from James Hogg iCapture Centre) for their various contributions.