Towards better understanding of complex biology: Ontology merging across data sources using multiple thesauri in semantic networks

Erich A. Gombocz[1]), Toshiro Nishimura[1]), Chuck Rockey[1])

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

Summary

Trying to understand the underlying complex biological functions represents one of the biggest challenges scientist face when dealing with experimental measurements from drug- or disease-related responses. To account for multiple effects resulting in the observed response, such an effort requires integration of data across dissimilar nomenclatures, taxonomies and ontologies and bridging the many differences in data formats towards a merged semantic knowledgebase.

This study presents an example of how to use one or multiple thesauri in conjunction with manual and automated mapping to obtain a common, coherent ontology with controlled vocabularies for both, data classes and their relationships. A "smart" merging must also account for hierarchical adjustments into super- or subclasses of the new, merged ontology tree whenever needed - specifically when integrating internal experiments with public domain reference sources with very specific taxonomies and hierarchies.

On examples of biomarker discovery and qualification, the application of such a technology across data from different disciplines and the merge of multi-OMICs experimental findings with public sources, the pros and cons of using formal semantic formats and data models (RDF/N3/OBO/OWL) meaningfully together with output from queries to relational database are addressed. The use of a common semantic network approach towards understanding of biological responses is exemplified by its application to toxic perturbations involving feedback from multiple pathways which - without coherent merging- would not be detectable. Its general applicability for a biology-based biomarker diagnostics in personalized medicine, patient stratification for disease profiling or toxicity risk assessment in rational drug design are evaluated.

Challenges

- Data coherence: different taxonomies, ontologies, non-standardized vocabularies.
- Ontology merging requires tree-level propagation adjustments for super-classes and sub-classes
- In many cases, data relationships are not a priori contained in the datasets, requiring the ability to add or map missing relationships.
- Multiple, incompatible semantic standards make merging towards a common ontology extremely difficult; this also impacts applicability of query and reasoning.
- Data type-selective approaches and pre-defined taxonomies are not generally applicable across experimental and public domain data resources.
- Lack of intuitive, science-driven tools makes network analysis unattractive to most researchers.
- Scalability and performance issues confine most network approaches to relatively small datasets.
- Biological understanding of complex processes is in its infancy.

Methodology

- Use import mapping to consolidate relationships; separate numeric from non-numeric content during class generation to account for scaling using numerical properties.
- Apply controlled vocabularies when importing or merging, using one or multiple thesauri
- Provide thesaurus authoring capabilities for both, data classes and their relationships.
- Allow for proper handling of non well-formed RDF.
- Use a parsing methodology which can handle different data representations.
- Enable import from or directly connect to via URIs.
- Exercise a multi-functional query tool supporting graphical and SPARQL queries to reduce network complexity.
- Facilitate output of the resulting knowledgebase with or without its graphical representation in multiple semantic formats (RDF, N3 or triples store(s) backend).

Approach

- For output from queries to relational databases (in tsv, csv, xml, fqml) or from devices or databases using PSI-MI, an interactive graphical/textual import mapper for relationships is applied; such mapping also can be configured for automatic background mapping when deployed in conjunction with web-based query to multiple relational databases.
- Use, author and apply one or multiple thesauri for consolidation of classes and relationships
- "Drag-to-knowledgebase" function with automatic classification of semantic data in RDF, N3, OBO and OWL (local or remote file system)
- Permit direct URI input to web-based reference resources
- Utilize entity browser with sorting (by subject, predicate and object) and paging for large datasets to navigate through instance properties
- Provide means to reduce network complexity: apply criteria for connection depth, numeric scaling and weighting and other conditions (such as at least, at most, exactly n connections)
- Employ a built-in query tool which supports graphical, textual and SPARQL queries to specify conditions, and then re-plot the resulting sub-networks to reveal hidden or unknown functional biological relationships and their entities in context.

Results

- This poster demonstrates a semantic approach which overcomes most obstacles for meaningful data integration by applying import mapping, one or multiple thesauri for handling of synonyms or naming inconsistencies and ontology merging across their classes and hierarchies.
- Being able to take advantage of existing ontologies (such as GO) , the analysis of the resulting network of multi-OMICs experiments combined with brought-in public reference resources and pathway information enabled us to visualize, query and qualify potential toxicity biomarker panels with regard to their systems biological relevancy.
- In several cases, we were not only able to validate changes in biomarkers with common mechanisms of toxicity previously described (oxidative stress, liver function, Ketoacidosis), but additionally were able to discover new pharmacodynamically and biologically linked components for further investigation of their functions within the biological system.
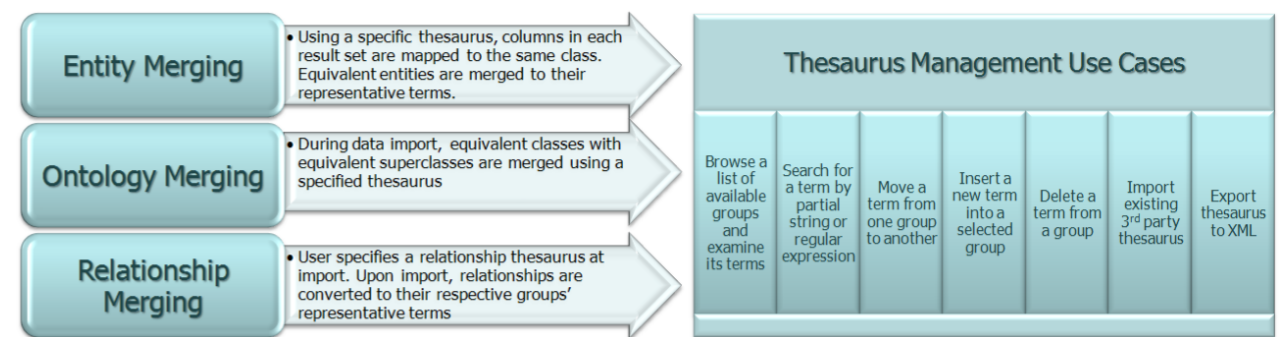

Fig.1: Overview of requirements for ontology merging (left) and use cases for Thesaurus management functions
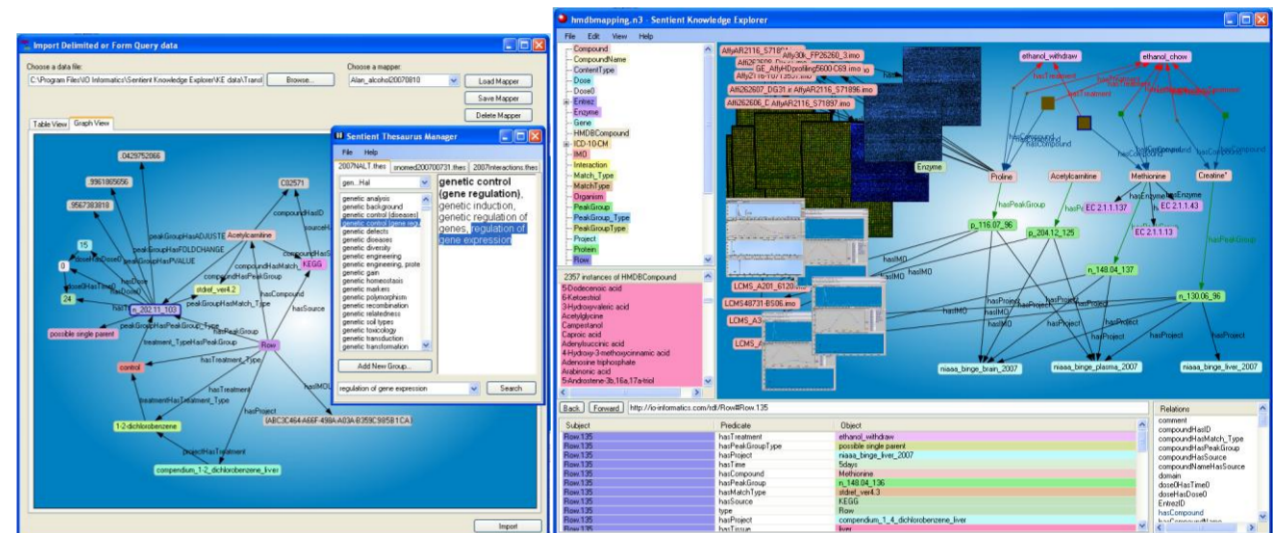

Fig.2: Mapping and managing nomenclature via Sentient Thesaurus Manager™(left); a common knowledgebase from genes, metabolites, their raw data references and public resources in Sentient Knowledge Explorer™ (right)
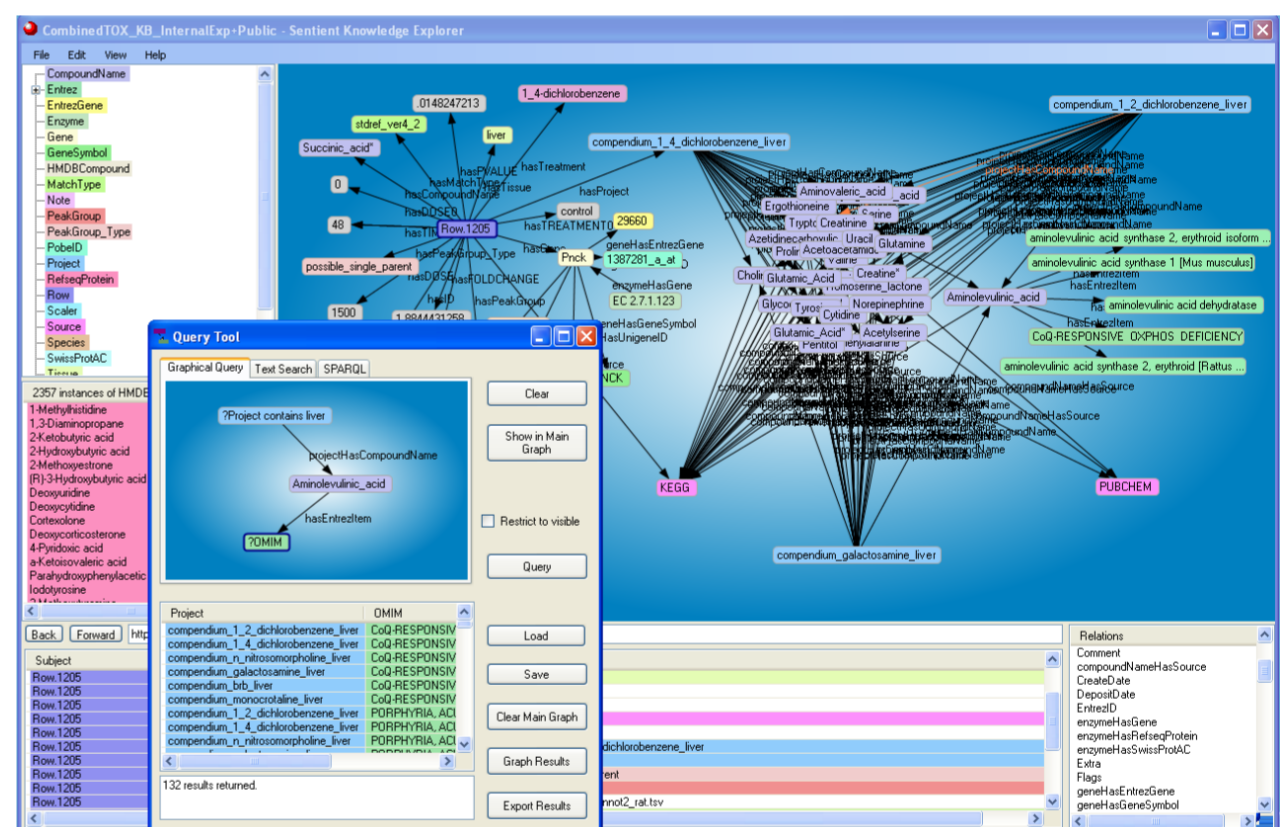

Fig.3: Biomarker discovery, query and validation: Direct drill-out and query from experimentally observed toxicity responses mapped to multiple public reference resources (NCBI, UniProt, IntAct, BioGrid, KEGG, HMDB)

Conclusions

The presented semantic approach for coherent data integration provides the foundation for discovery, qualification and validation of biomarkers in a systems biology context. Specifically, we were able to:

- Coherently merge of experimental and public data across standards, taxonomies and ontologies into an integrated, semantically organized data and relationship network
- Focus on visualization and analysis of a sub-network of specific scientific interest (e.g. biomarker qualification across tissues, different treatments, different genetic responder profiles, etc.) in a complex, large dataset
- Correlate metabolites across different treatments to discover common effects for a class of drugs
- Qualify potential biomarker panels for toxicity across tissues and toxicants
- Gain insights in complex biological functions involving multiple pathways through sub-network analysis.

Future

- Generation of a foundation for reasoning across complex translational research data.
- Application to personalized medicine (patient stratification, pre-symptomatic disease detection, responder profiles based on genotypes, drug interaction profiles).