

Applied Semantic Knowledgebases (ASK®): Changing how knowledge is built and applied in Life Sciences and personalized medicine

Robert Stanley¹, Zack Rhoades¹, Erich Gombocz¹

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

Summary

While predictive biology has been a major goal in the development of safer drugs and more effective therapies in the clinics for quite a while, its promises have – despite major advances in analytics – been challenged by 2 major factors: the difficulties of meaningful semantic integration of heterogeneous experimental and public data, and the complexity in understanding the biological functions involved.

Building on advanced data access and integration capabilities, IO Informatics Sentient framework uses semantic patterns to create predictive network models using virtually any combination of internal experimental data and / or external published information. These patterns apply semantic SPARQL query technology to build complex searches across multiple information sets. SPARQL is a semantic search technology capable of detecting patterns within and between different data types and relationships, even if the initial datasets are not formally joined under any common database schema or data federation method. Such patterns or models are then placed in an Applied Semantic Knowledgebase (ASK) which is unique to a specific research focus, providing a collection of specific models applicable to screening and decision making. Applications include target profile creation and validation, compound efficacy and promiscuity screening, toxicity profiling and detection, disease signatures, predictive clinical trials pre-screening, and patient stratification.

Specifically, in biomarker discovery generating test models which can be qualified, refined and validated for broader use has been previously challenging and required considerable efforts.

The poster describes what is involved in using this new technique, how it changes the field by allowing researchers to better understand how the organism responds to experiment-induced system changes and to recognize mechanistic aspects of biomarkers at a functional level, and, how it is applied to predictive screening based on multi-modal datasets.

It demonstrates, that ASK makes it possible to actively screen previously disconnected, distributed datasets, to identify and stratify results – delivering applications used for decision making in life science and personalized medicine, and changing the way, how knowledge is built and applied.

Challenges

- Data coherence: different data sources, taxonomies, ontologies, non-standardized vocabularies.
- Complexity of network analysis in general and lack of intuitive, science-driven tools makes such approaches non-appealing to researchers and industry.
- Multi-OMICS expression changes can represent very different biological processes and typically exhibit sets of multiple overlapping alterations.
- Modeling biological systems to predict phenotypic outcome is still very incomplete.
- Validation of classifiers and closeness of fit between unknowns and model are demanding.

Methodology

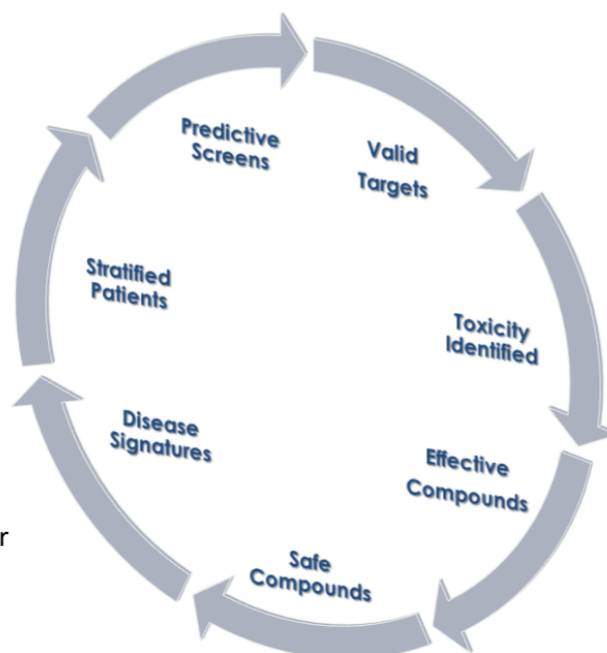
- Identify statistically significant perturbation in multiple modalities with robust correlation between independent analytical results via query.
- Merge and map results into a semantic framework to visualize, investigate and analyze data relationships.
- Associate significant elements of those networks with reference data sources, using thesauri to consolidate data class and relationship synonyms, and combine experimental data with literature
- Scale potential markers using numerical properties to reduce network complexity and pre-select classifiers.
- Save the resulting sub-network as SPARQL query, and represent your model as an array of such queries.
- Refine model with test cases, then apply to unknowns for screening and use the confidence of the match (“hits-to-fit”) for informed decision-making.

Results

- This poster demonstrates how data from diverse experimental and public resources are queried and merged into a semantic framework to provide insights into complex biological functions by putting data in context of their relationship to others.
- Network queries are directly generated from interactive, user-selected nodes on the network, requiring no knowledge of SPARQL query language.
- Sets of such SPARQL queries are captured and saved in arrays representative for a specific biological function and have been applied in decision-support in compound toxicity studies for Hepatotoxicity and nephrotoxicity and for predictive patient screening for organ failure or acute organ rejection.

Discussion

- As Applied Semantic Knowledgebases (ASK™) represent a novel approach towards complex biological responses, the qualification criteria to select classifiers and the algorithms involved in the statistical approach are crucial.
- Semantic integration and merging of data assures coherence and provides a solid base to relevant network analysis.
- Being able to create complex models in an easy, automated way, makes it universally applicable.
- By providing an array of network-based models, a high degree of confidence can be obtained – particularly, if responses are accompanied by their closeness of fit to qualify the prediction.



- While this concept already is actively applied in a wide area of interests in pharmaceutical research, life sciences and personalized medicine, its function as knowledge application to provide decision support ranges from targets to compounds to patient treatment and screening.

Figures

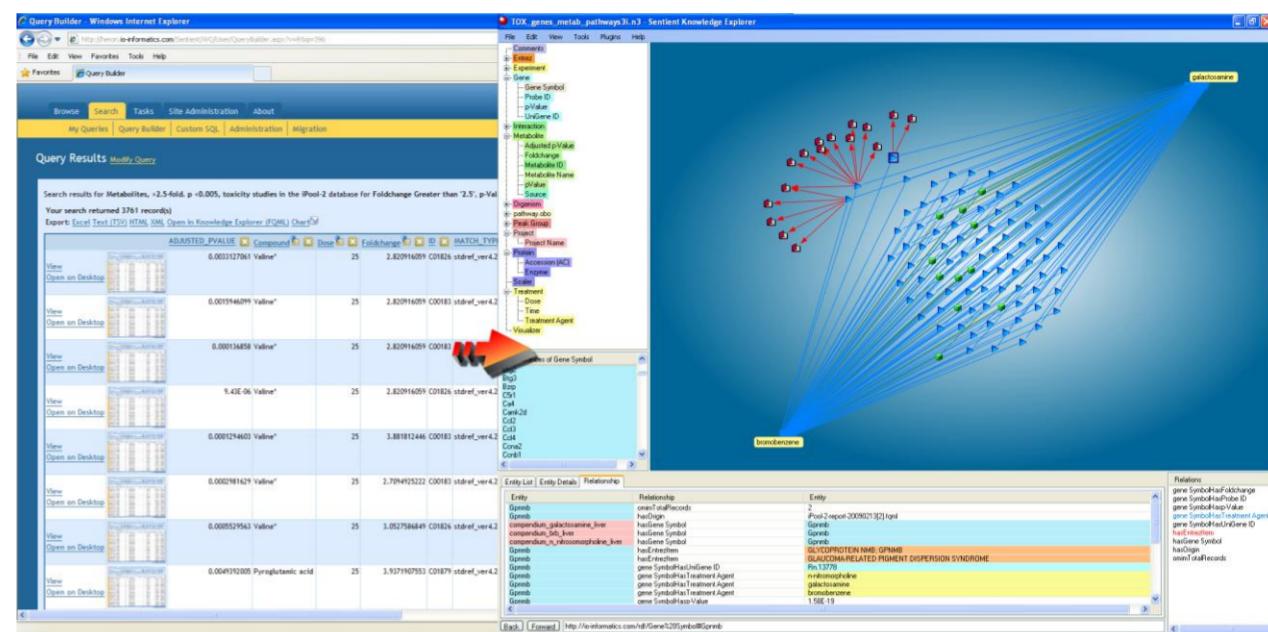


Fig. 1: Semantic data merging: Query results from multiple sources are represented in a common ontology network

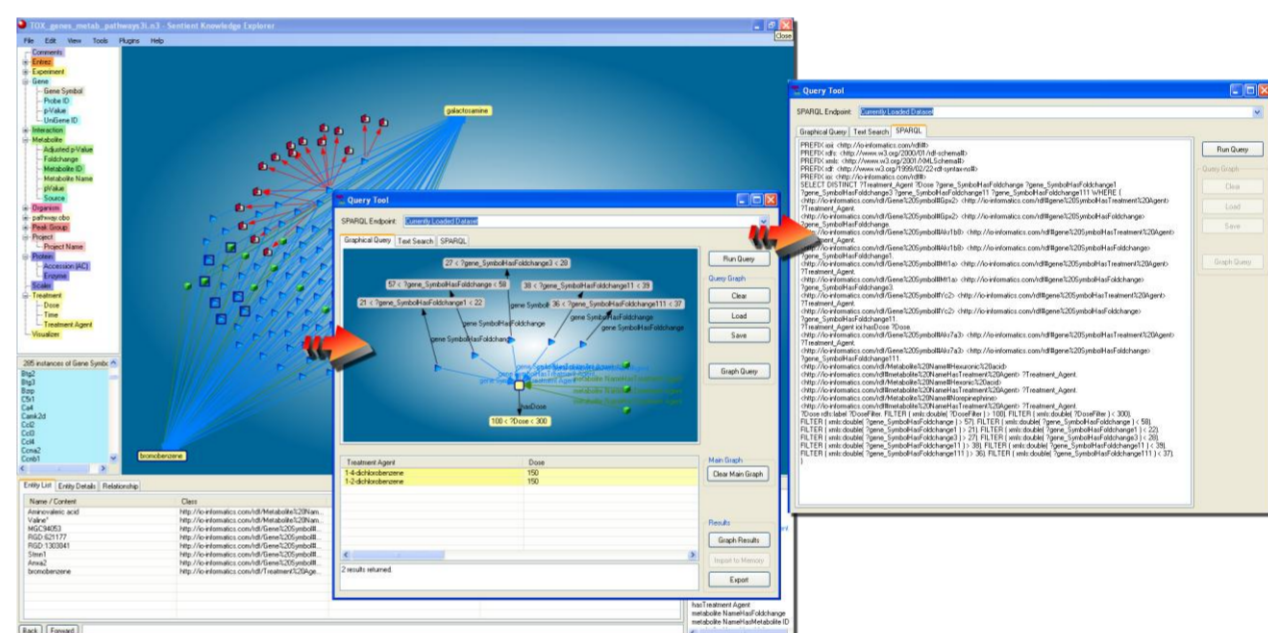


Fig. 2: SPARQL creation directly off the graph: Selecting nodes generates graphical query and SPARQL representation

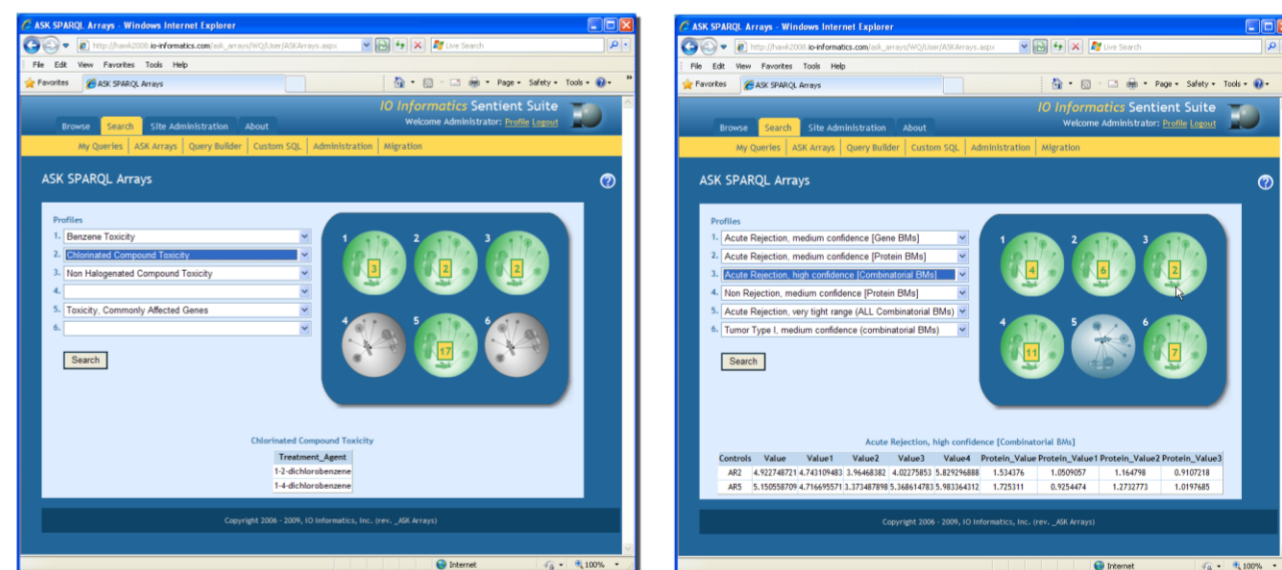


Fig. 3: ASK arrays on the web: Predictive screening for decision-making (left Toxicity – right Organ failures)

Conclusions

- Applied Semantic Knowledgebases (ASK™) represent a novel way to tackle the inherent complexity of biological responses. Using arrays of semantically-based models in an easy-to-use fashion accounts for its appeal to researchers in life sciences and personalized medicine, who are faced with complex biological questions and rely on decision-support day-by-day.
- The ability to use, share and apply knowledge based on sophisticated network models via an intuitive web tool hiding the underlying complexity from the user and rather providing concise information which data (targets, compounds, diseases, patients) fit the model, and how good the match is in each particular case, is changing the way how knowledge is built, refined and applied in life sciences and medicine.

Acknowledgements

Some data examples are part of a study performed under the NIST Advanced Technology Program (ATP), Award # 70NANB2H3009. This work was also made possible through insights and many discussions in IO Informatics' Working Groups on "Semantics in Life Science" and "Informatics for Personalized Medicine". The authors would like to acknowledge the Working Group members Jonas Almeida, Alan Higgins, Pat Hurban, Bruce McManus, Ted Slater, Mark Wilkinson, Uwe Christians, Jack Collins, Dan Crowther, Amar Das, Herb Fritsche, Kathy Gibson, William Hayden and David Stanley for their various contributions.