

Realizing personalized medicine with semantic technology: Applied Semantic Knowledgebases (ASK[®]) at work

Robert Stanley¹⁾, Erich Gombocz¹⁾, Zack Rhoades¹⁾

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

Summary

Building on core data access and integration capabilities, Sentient software applies semantic patterns to create predictive network models using virtually any combination of internal experimental data and / or external published information. These patterns apply extended semantic “Visual SPARQL” query technology to build complex searches across multiple information sets. SPARQL is 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 are then placed in an Applied Semantic Knowledgebase (ASK) which is unique to a specific research focus, providing a collection of applicable to screening and decision making. Applications include hypothesis visualization, testing and refinement; target profile creation and validation; compound efficacy and promiscuity screening; toxicity profiling and detection; disease signatures; predictive clinical trials pre-screening; and patient stratification.

Using a customer example for application of this technology to personalized medicine - for presymptomatic detection, scoring and stratification of patients at risk of organ failure according to combined genotypic and phenotypic information – the power of ASK is demonstrated. Insights gained from semantically joining coherent findings despite their different methodologies allow researchers to better understand mechanistic aspects of biomarkers for organ failure at a functional level; and to apply complex screening algorithms using SPARQL and connected statistical methods for highly sensitive and specific patient stratification.

Using ASK makes it possible to actively screen previously disconnected, distributed datasets, to identify and stratify results - delivering applications to be used for decision making in the life science industry and in personalized medicine.

Challenges

- Multi-OMICS expression changes – despite all resulting from the same disease state – represent very different biological processes and can exhibit the sum of multiple overlapping alterations from several pathways.
- Biological system’s networks and pharmacodynamic data network correlations are not necessarily functionally connected.
- Different semantic standards in laboratory and clinics make merging towards a common ontology extremely difficult, which impacts also querying and reasoning.
- Relationship consolidation and class hierarchy adjustment may be required to make inference and reasoning meaningful
- Effective, intuitive SPARQL-based query tools for model building, refinement and applying as knowledge profiles have been missing.

Methodology

- Identify statistically significant perturbation in multiple modalities with robust correlation between independent analytical results.
- Merge and map results into a semantic framework to visualize, investigate and analyze data relationships.
- Associate significant network elements 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.
- Weigh markers according to mechanistic insights and biological relevancy to assure the network makes sense from a biological systems perspective
- Save the resulting sub-network as SPARQL query, and represent the model as array of such queries.
- Refine model with test cases, then apply it to unknowns for screening. Use scoring to represent the confidence in the match (“hit-to-fit”) for informed decision-making. Re-evaluate and refine the model with newly confirmed cases to increase prediction precision.

Results

- Diverse experimental and public resources were merged into a common semantic framework; the data networks were enriched with knowledge networks to provide mechanistic insights into complex biological functions affecting disease state and progression.
- SPARQL graph queries to qualify biomarker classifiers are directly generated from interactive, user-selected network nodes without requiring 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 predictive patient screening for organ failure or acute organ rejection.

Discussion

- Applied Semantic Knowledgebases (ASK[™]) represent a novel approach towards complex biological responses. Therefore, the qualification criteria to select classifiers and the algorithms involved in the approach are crucial.
- Semantic integration and merging of data assures coherence and provides a solid base to relevant network analysis, allowing to incorporate mechanistic aspects of biological functions.
- 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 – specifically, 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 closes the loop from targets to compounds to patient treatment and screening back to drug discovery.

Impact

- Cardiovascular disease (heart disease, diseases of the blood vessels and stroke) accounts for the death of more 30% of the Western population. In 2006, there were 33,832 Canadians on renal replacement therapy, and this number is expected to double over the next 10 years.
- Tissue biopsies are currently the only way to monitor transplant patients for organ rejection. They are invasive but necessary to fine-tune the dosage of immunosuppressive drugs required by every transplant patient. Too small a dosage can result in organ rejection and potential organ failure; too much leaves patients susceptible to dangerous infections and cancer. Biopsies are costly, too: heart transplant patients undergo at least a dozen biopsies in the first year after transplant, at a cost of \$5,000-\$10,000 each. The ability to predict and diagnose rejection of transplanted organs with a simple, inexpensive blood test significantly reduces the need for biopsies and the burden on the healthcare system.

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.

Figures

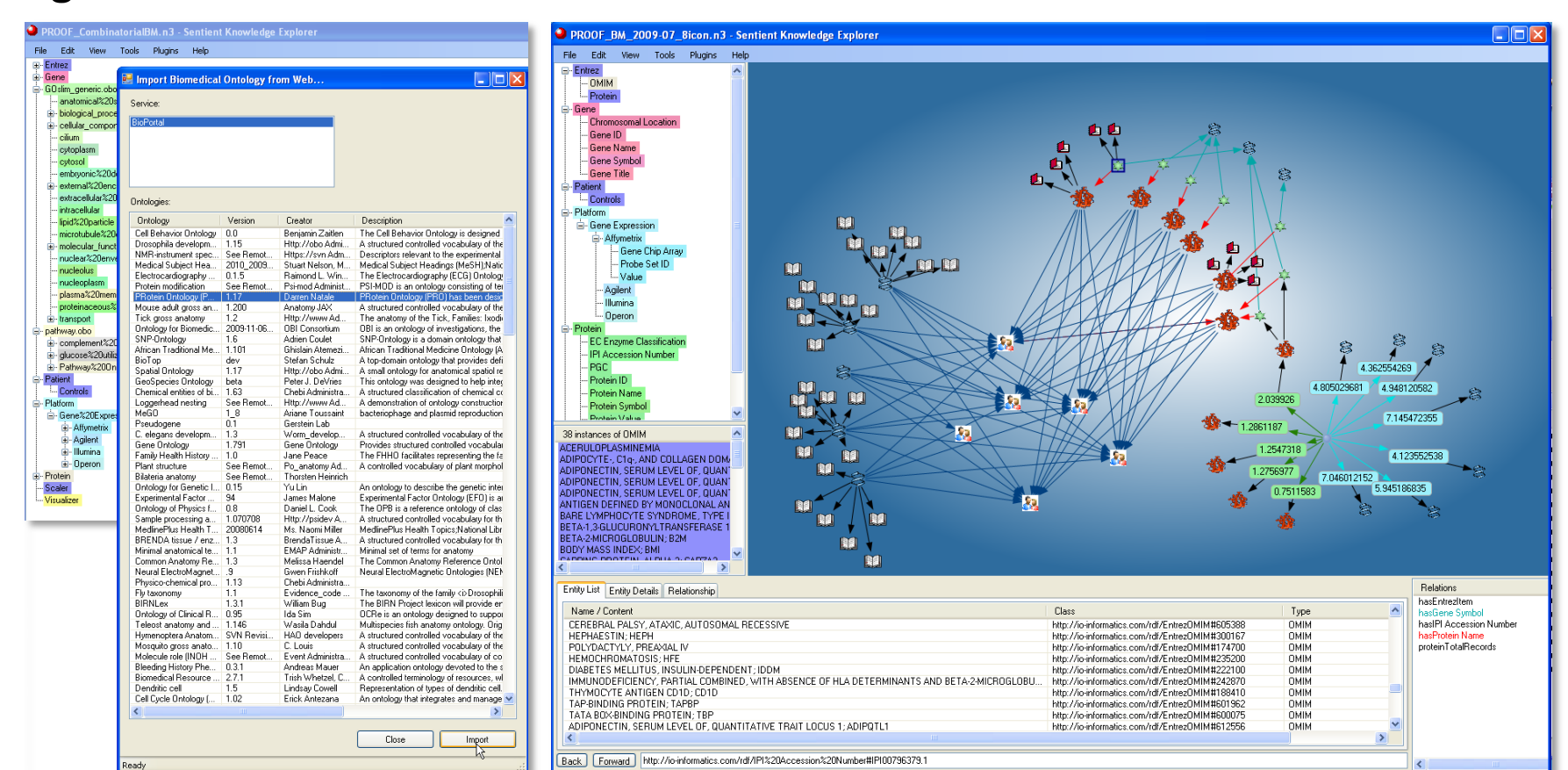


Fig. 1: Semantic data merging: Ontology import (left) and public reference-enhanced experiments in a common ontology network (right)

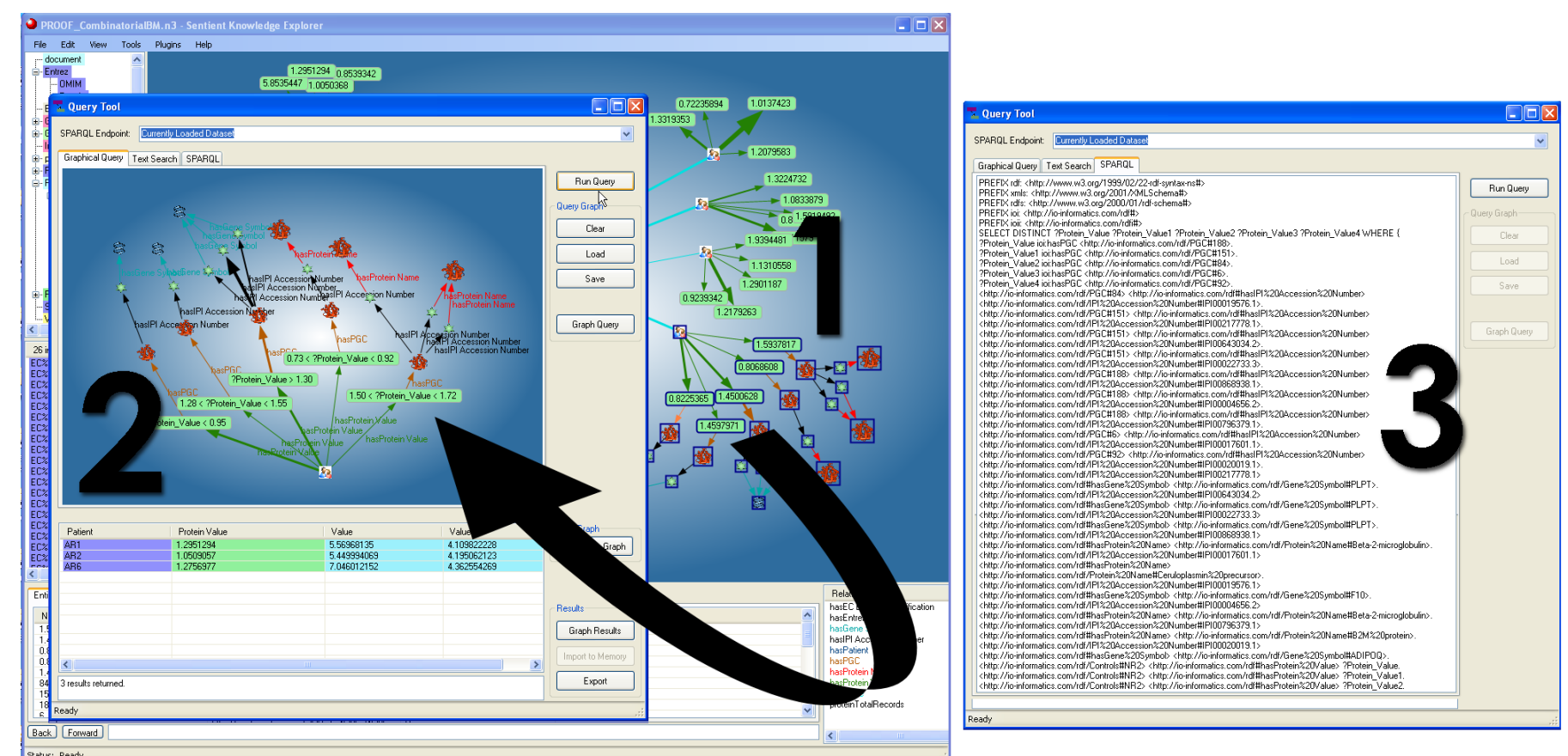


Fig. 2: SPARQL creation directly from graph: Selecting nodes from the main network graph (1) generates a visual SPARQL representation of the query(2) and the actual SPARQL statement (3) automatically

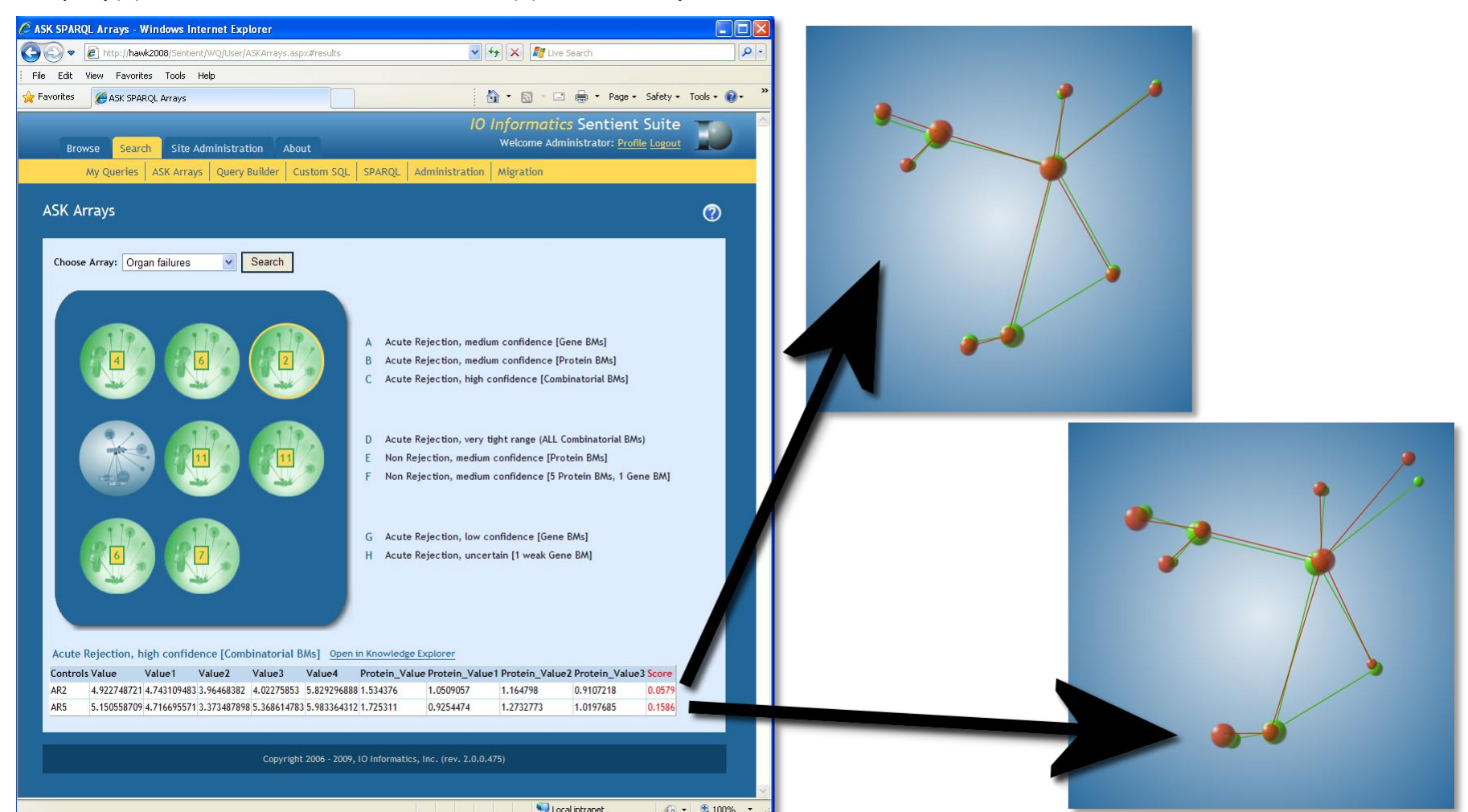


Fig. 3: Web-browser accessible ASK arrays: Predictive screening as decision-support for organ failures (left: main interface with scoring) and “hit-to-fit” representation as decision support tool for goodness of prediction (right)

Acknowledgements

This work was supported through valuable suggestions and 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 and David Stanley for their various contributions.