# Applied Semantic Knowledgebases (ASK®): Confident decisions in personalized medicine using semantic biology models

Erich Gombocz<sup>1)</sup>, Robert Stanley<sup>1)</sup>, Chuck Rockey<sup>1)</sup>, Zack Rhoades<sup>1)</sup>, Jason Eshleman<sup>1)</sup>

<sup>1</sup>IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, U.S.A. [Correspondence: egombocz@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 been challenged by the difficulties of meaningful semantic integration of heterogeneous experimental and public data, and the complexity in understanding the involved biological functions.

Building on core data access and integration capabilities, Sentient software applies semantic patterns to create predictive biology 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 federation method. An Applied Semantic Knowledgebase (ASK) unique to a specific research focus contains arrays of such pattern, providing a collection 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.

On customer examples of applications in personalized medicine - for presymptomatic detection, scoring and stratification of patients at risk of organ failure according to combined genotypic and phenotypic information; understanding the influence of inflammation at the onset of plaque rupture; and, effectiveness assessment of combinatorial treatments in cancer therapies — the power of ASK is demonstrated. Insights gained from semantically joining experimental correlations despite different methodologies and enriching them with curated public domain knowledge networks allow researchers to better understand mechanistic aspects of biomarkers 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 used in decision making in life science and personalized medicine.

#### 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 represent very different biological processes and exhibit sets of multiple overlapping alterations.
- Validation of classifiers and closeness of fit between unknowns and model are demanding.

# Methodology

- Identify statistically significant changes 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 each model with test cases, apply it to unknowns for screening and use scoring of the match ("hits-to-fit") for informed decision-making with high confidence.

#### **Results & Discussion**

- This poster demonstrates an innovative approach towards the integration of experimental and public data sources, the visual exploration of this integrated environment, and the building of network patterns characteristic for specific biological functions.
- Semantic integration and merging of data assures coherence, harmonizes synonyms and different terminologies, and provides a consistent base to relevant network analysis.
- SPARQL queries are captured and saved in arrays representative for a specific biological function.
   Being able to create complex models in an easy, automated way makes the method universally applicable.
- By providing an array of network-based models, a high degree of confidence can be obtained even more so, if responses are accompanied by their closeness of fit to qualify the prediction.
- As Applied Semantic Knowledgebases (ASK™) represent a novel approach towards complex biological responses, qualification criteria to select classifiers and the algorithms involved in the statistical approach are crucial.
- ASK has been actively applied as decision-support in predictive pre-symptomatic patient screening for organ failure (PROOF<sup>2</sup>), identification of plaque stage prior to rupture in high risk plaque (BG Medicine HRP<sup>3</sup>), and the comparative effectiveness of cancer treatment for prostate cancer (PCC<sup>4</sup>).
- While this technique already is applied in several areas 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.

#### **Impact**

- Applied Semantic Knowledgebases (ASK™) provide researchers in life sciences and clinics, who
  are faced with complex biological questions and rely daily on decision-support, with a novel,
  directly applicable approach to address their needs.
- Being able to use, share and apply knowledge based on sophisticated network models via an
  intuitive web tool which hides the underlying complexity from the user, yet provides 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 personalized medicine.

#### **Figures**

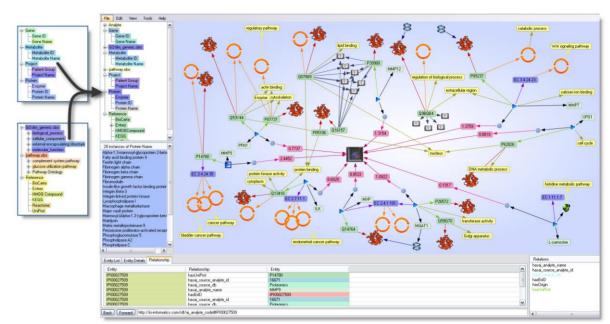


Fig. 1: Semantic data merging of experimental correlation networks and knowledge networks: Ontology import (left) and public reference-enhanced experiments in a common ontology (right)

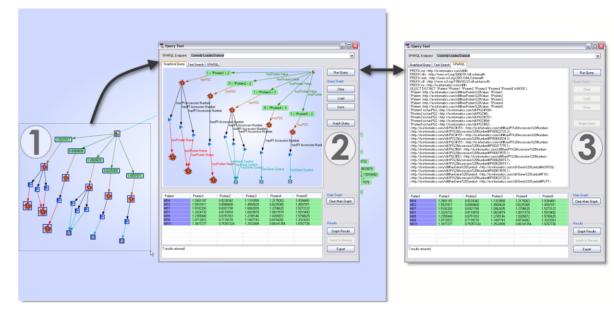


Fig. 2: SPARQL creation from graph: Node selection in the main network (1) generates a visual SPARQL representation of the query(2) and the actual SPARQL statement (3) automatically

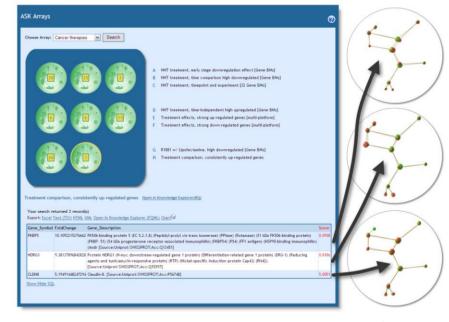


Fig. 3: Web-browser accessible ASK arrays: Predictive screening as decision-support for comparison in effectiveness of treatment (left: main interface with scoring) and "hit-to-fit" representation for goodness of prediction (right)

### Conclusions

- Applied Semantic Knowledgebases (ASK™) represent a novel way to address the complexity of biological responses. Applying 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 confident decision-support day-by-day.
- Actionable inferences about disease states and treatments using combined clinical, genomic, proteomic, metabolic and molecular phenotypic data in conjunction with mechanistic insights from public knowledge networks via a web-based tool is a remarkable step in patient-centric personalized medicine.

# References

- 2) Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada.
- <sup>3</sup>) BG Medicine Consortium on High Risk Plaque (HRP), Waltham, MA, U.S.A.
- <sup>4</sup>) UBC Prostate Cancer Centre (PCC), Vancouver General Hospital, Vancouver, BC, Canada

## **Acknowledgements**

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