Practical use with major implications: Applied Semantic Knowledgebase (ASK™) for Predictive Biology

Erich Gombocz
VP, Chief Scientific Officer
IO Informatics, Inc.
Outline

• What is an “Applied Semantic Knowledgebase” (ASK™)?

• How is it done?
  • Building & exploring knowledge
  • Creating & applying SPARQL queries
  • Refining & qualifying models

• Scenarios & practical implications

**Live DEMO**
• Use case: Combinatorial biomarker for toxicity

• Summary & discussion
  • Impact of ASK
  • How far are we?
What is an ‘Applied Semantic Knowledgebase’?

Graph queries represent active knowledge contained in ASK.
Framework to integrate, unify and combine data for knowledge extraction.

Looking at data by how it relates to other data. Allows information to adapt and evolve.

ASK

SENTIENT

SEMANTICS

SPARQL
Steps to get there ...

1. **Unify & analyze data**
2. **Create semantic networks to visualize and explore biomarkers**
3. **Capture combinatorial marker patterns in ‘SPARQL Arrays’**
4. **Test hypotheses, qualify & validate model**
5. **Apply ASK for predictive biology**
Knowledge building and exploration

• Merge data dynamically into an extensible & reconfigurable ontology
• Apply thesauri for classes, entities and relationships

Entity Merging
• Using a specific thesaurus, columns in each result set are mapped to the same class. Equivalent entities are merged to their representative terms.

Ontology Merging
• During data import, equivalent classes with equivalent superclasses are merged using a specified thesaurus

Relationship Merging
• User specifies a relationship thesaurus at import. Upon import, relationships are converted to their respective groups' representative terms
Manage ontologies, vocabularies
Example Case: Toxicity compendium and Alcohol study

Objective: Better understanding of systems biology of toxicity

Gene expression [GEP] (Affymetrix MA)
Metabolic profiling [BCP] (Bruker Daltonics LC/MS)
Quantitative tissue analysis [QTA] (BiolImagene)
Metadata from internal LIMS

Set of known toxicants (hepatotoxicants)
Multiple tissues, different experimental animal models (rats)

Enrich experiments with public knowledge (NCBI, HMDB, KEGG, IntAct, BioGRID, PubMed …) for causal reasoning on the biology (pathways, biological functions)

Goal: Combinatorial biomarkers for toxicity prediction
- Identify system perturbation
- Commonly affected genes and metabolites
- Tissue-specific vs. non-tissue-specific effects
- Compliment & qualify experimental results
- Establish a sub-network of interest
Create SPARQL queries

- Automatic SPARQL directly from graph
- Choose the sub graph you are interested to explore
- Set confidence ranges for numeric values
- Run query on training set for iterative refinements
- Focus on science, not on manual query editing
- Generate SPARQL queries “without SPARQL”
Hide complexity! *(unless needed)*

- SPARQL generation occurs behind the scene
- Scientists don’t need to know anything about SPARQL
- BUT if you need, you can edit, cut, copy, paste …
- Allows those who are familiar with SPARQL to review and modify directly for testing
- ASK contains collections of such queries
Test patterns for qualification

Immediate value from ASK SPARQL queries

Example:
- find compounds likely to exhibit a specific type of toxicity
  - Step 1: Load Toxicity ASK
  - Step 2: Run query
  - Step 3: Review results

Result:
- 3 compounds returned

✓ Model Tested
Apply ASK SPARQL arrays to screen

- Apply ASK SPARQL array queries via web
- Example:
  - 3 different toxicity types
  - Search for matches
- Result:
  - All 3 profiles found matches
  - 3 compounds for Benzene Toxicity (1) are listed for further analysis

✓ Easy and rapid screening
Practical implications

- Sets of models can be easily derived from graph and aggregated in ASK for decision support

- ASK patterns are directly applied to screening

- Applications include
  - Drug target profiles
  - Compound efficacy screening
  - Toxicity profiling and detection
  - Disease signatures
  - Patient selection for clinical trials
  - Patient stratification

- ASK is effective in organizing and applying the knowledge contained in your data

- **ASK gives you a decisive competitive advantage in biomarker-based predictive biology**
Live DEMO
What did we learn?

• Semantic data integration of experimental and public data provides the framework for meaningful biological model generation

• Sub-network visualization (intersections, exclusions) is directly transformed into SPARQL queries

• SPARQL queries describe network-derived models

• Arrays of SPARQL queries contained in ASK allow to screen for complex biological functions
Take-home message

• Semantic data integration puts multi-modal experimental and public data in context

• Network exploration facilitates capturing marker classifiers

• Query pattern are directly derived from graph

• Confidence in model can be iteratively refined

• ASK patterns are directly applied to screening of unknown datasets

Result:
‘ASK’ excels in testing & applying complex biological models
Acknowledgements

**NIST Advanced Technology Program (ATP)**
Award # 70NANB2H3009
Frank Barros

**CLDA / Cogenics**
Tom Colatsky, Pat Hurban, Imran Shah, Hongkang Mei, Alan Higgins, Maureen McBride-Simon

**IOI Semantic Working Group**
Jonas Almeida, Alan Higgins, Pat Hurban, Bruce McManus, Ted Slater, Mark Wilkinson

**Semantic standards development**
CSHALS, W3C HCLSIG, Oracle 11G Test Group
Discussion

Erich Gombocz
egombocz@io-informatics.com

IO Informatics, Inc.
2550 Ninth Street, Suite 114
Berkeley, CA 94710-2552
U.S.A.

Phone  (+1) 510-705-8470
Fax    (+1) 510-705-8476
www.io-informatics.com