Computational R&D in Action: Integrating Correlation and Knowledge Networks For Treatment Response Modeling and Decision Support

Erich Gombocz[1)], Robert Stanley[1)] , Jason Eshleman[1)]

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

Summary

With computational R&D spanning multiple information domains - dynamic, flexible and extensible solutions are required to cope with the demand for knowledge management and data sharing. In addition to privacy and security concerns, analytical and demographic data from instruments, files, images, databases and the web need to be put together in the context of biological mechanisms and clinical relevance in order to generate applicable knowledge. The integration challenges become even more apparent when addressing complex medical data from hospitals, research laboratories, external facilities and public resources. Such considerations require a new approach in order to harvest data relationships and interactions in a meaningful way to generate valuable knowledge. This approach creates applicable models for effective decision making in discovery, development and personalized medicine applications.

This poster demonstrates how semantic data integration, where data are placed in their functional context with other data, changes the playing field. Semantics harmonize synonyms and nomenclature to account for common terminology and data relationships across life sciences research and clinical domains. Additionally, semantics allow for relationship mining, inference and reasoning in a systems approach. Dynamic, flexible resource description framework (RDF) databases, ubiquitous SPARQL endpoints and "linked open data" initiatives are becoming important tools and are keys for research demanding global collaboration. Graph representations of merged data, including correlation networks and formally mechanistic reference networks, allow for identifying target classifiers in discovery and focusing on relevant dimension-reduced subnets. Furthermore, knowledge from powerful systems-biology based models foster an understanding of complex biological processes (e.g., diseases & disease states, pre-disposition to certain biological responses, patient stratification for trials and treatment, predictive risk assessment for tumor growth, organ rejection, toxicity assessment and severity of drug side effects).

The poster presents several customer cases – utilizing reasoning for excipient influence on drug formulation stability despite imprecise connections; comparative assessment of effectiveness of combination prostate cancer treatment; and combinatorial biomarker-based toxicity classification. These examples demonstrate capabilities gained from semantic integration of analytical correlation networks with public knowledge networks towards a better understanding of complex biologic interactions relevant to personalized medicine. This technology is actively applied in production environments for 1.) combining and reasoning across experimental and published data to create target profiles identifying effective compounds and warning of toxic indication; and 2.) creating actionable inferences and intelligent screening patterns about disease states and treatments. These methods combine clinical, genomic and molecular phenotypic data within Applied Semantic Knowledgebases (ASK™) to assert optimum pharmaceutical development towards patient-centric personalized medicine.

The implications of using Applied Semantic Knowledgebases (ASK™) in drug development, personalized therapy, improvement of quality of life for patients and its impact on pharmaceutical development and healthcare in general are demonstrated.

Challenges

- Most relational data sources have isolated structures and schemas, defined for narrow contexts.
- Dynamic data require flexible integration structures capable of cross-boundary contextualization.
- Research requirements, pharmaceutical production and post-market clinical routine data assessment are not aligned.
- Biomarker discovery often looks only on statistical correlation, overlooking if those also make sense biologically.
- Methods are cumbersome, tedious and time consuming with often poor results in usability.

Methodology

- Integrate experimental data (chemistry, OMICs, tissue images, clinical assays) to generate a semantic correlation network in an extensible ontology, using thesauri and class unification.
- Map and merge public data resources (including canonical pathways and functional biosystems) dynamically for ad-hoc mechanistic enrichment and knowledge expansion using formal concept.
- Merge ontologies (or parts of) and add data classes, instances and relationships via direct SPARQL endpoint queries selectively without the need to incorporate entire database content.
- Add information from drill-out searches to semantic RDF-based (directly) or relational (via D2R or similar) public databases where no SPARQL endpoints are available.
- Enrich the data network semantically by merging additional public knowledge resources to create better understanding from functional connections between the experimental correlation network and knowledge networks to validate the underlying biology.
- Generate a biomarker model from visual SPARQL arrays, assemble them into an Applied Semantic Knowledgebase for a specific biological problem (e.g. drug target response at a defined disease stage for a defined patient profile) for further, testing, validation and use.
- Use the Applied Semantic Knowledgebase for hypothesis testing, screening and decision-support.

Results

- This poster demonstrates how data from diverse experimental and public resources are merged into a coherent semantic framework for intuitive model building.
- Dynamic, flexible resource description framework (RDF) databases, ubiquitous SPARQL endpoints and "linked open data" provide rapidly deployed insights into complex biological functions through contextualizing of all available information.
- To establish biological models, generate network queries interactively from user-selected network nodes, without requiring knowledge of the SPARQL query language.
- Sets of such SPARQL queries are captured and saved in arrays representative for a specific biological function. Those arrays have been applied in decision-support in drug development and personalized medicine for predictive patient screening and decision support on treatment options.

Discussion

- Semantic integration and merging of all available resources (experimental correlations from treatments, prior knowledge from clinical trials and drug interactions) assures coherence and provides a solid base to relevant network analysis.
- Applied Semantic Knowledgebases (ASK™) represent a novel approach towards complex biological responses; as such, qualification criteria to select classifiers and the algorithms for the statistical approach are crucial.
- 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 as responses are accompanied by their closeness of fit to qualify the prediction.
- Using a secure web portal access to visually screen patients with confidence removes informatics barriers in clinics and makes it widely accessible.
- While this concept is already actively applied, its growing adoption as a knowledge application for hypothesis testing and decision support promises to impact life science and healthcare industries, through more effective drug discovery and development and improved patient care.

Conclusions

- Combining and reasoning across experimental data enriched with public domain knowledge provides a new level of insights in complex biological responses. Applied Semantic Knowledgebases (ASK™) using arrays of network models via an easy-to-use web portal provide a confident decision support tool for complex biological questions.
- Being able to use, share and apply knowledge via an intuitive web tool hiding the underlying complexity from the user, but able to inform with concise scoring about the goodness of fit of the model to each patient's profile is changing the way how knowledge is built, refined and applied in life sciences, pharmaceutical research & development and medicine.

Figures

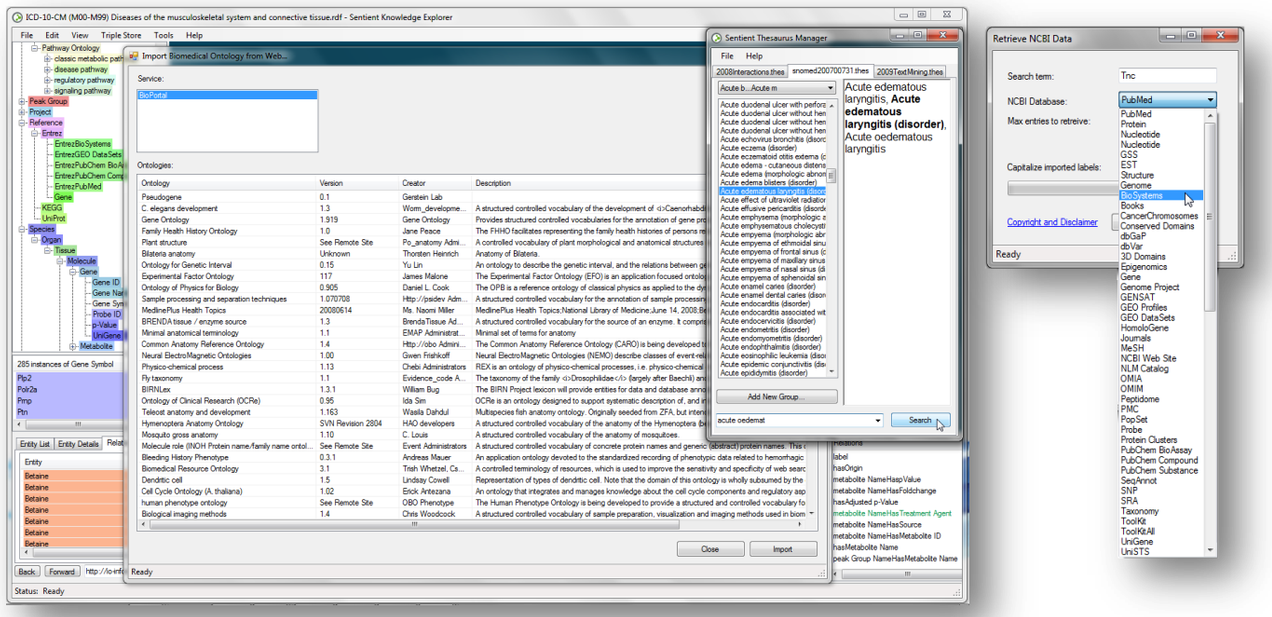

Fig.1: Integration of public ontologies, thesauri and database resources
Left: Merging ontologies (example: NCBO's BioPortal) and linked open data via SPARQL endpoints
Right: Adding mechanistic knowledge from public resources (example: NCBI Entrez) directly in RDF

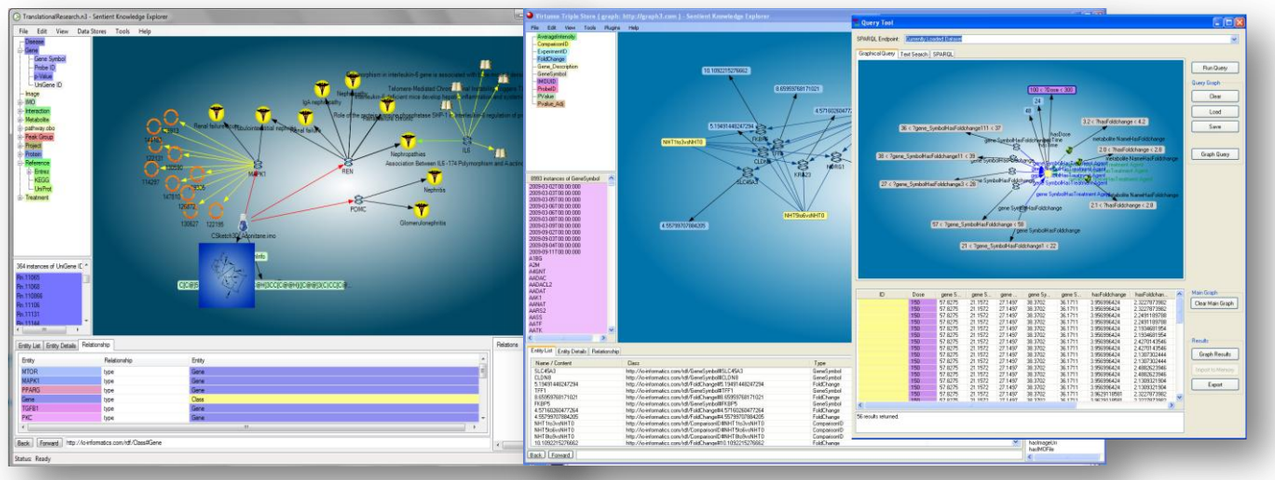

Fig.2: Semantic data merging
Main: Knowledge Explorer connecting multi-platform genomics (RNA, DNA, MAs) with public resources for external validation of disease-gene relationships in treatment effectiveness studies (Prostate Cancer)
Left overlay dialog: Visual SPARQL query generation to build, refine and qualify semantic model

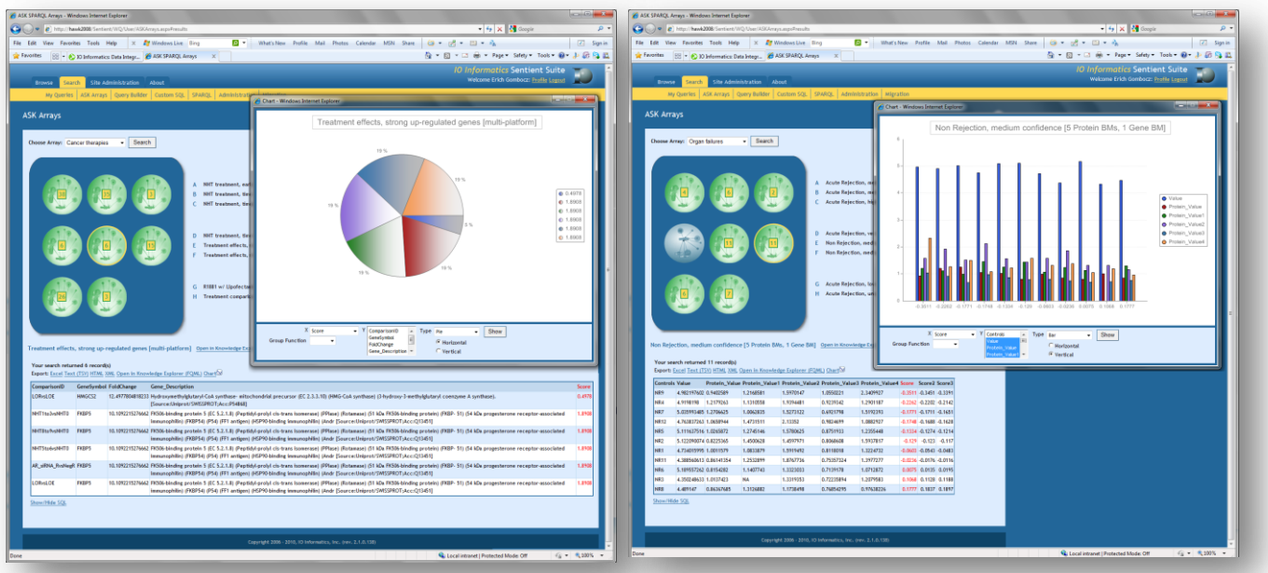

Fig.3: Actionable knowledge from integrated models
Left: Comparison of effectiveness of combinatorial prostate cancer treatments (genomic biomarkers)
Right: Systems model-based decision support for likelihood of organ rejection in heart transplantation (Combinatorial multi-OMIC biomarker panel)

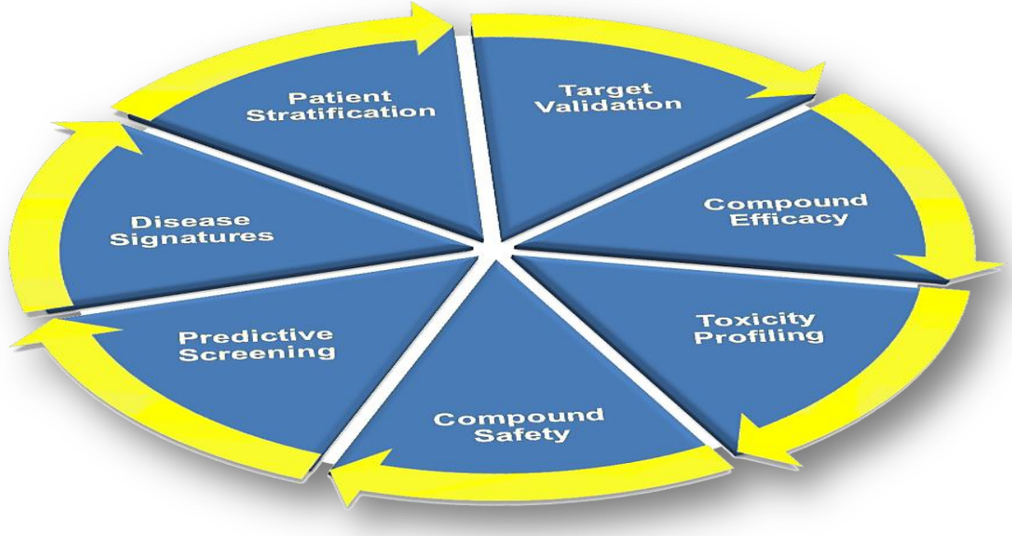

Fig.4: Applications Spectrum for Applied Semantic Knowledgebases (ASK™)

Acknowledgements

The authors thank **Colleen Nelson** (collaborating Prostate Cancer Centres) and **Bruce McManus** (PROOF and iCapture Centres) for their leading research and datasets. Many useful discussions in IO Informatics' Working Groups on **"Semantics in Life Science"** and **"Informatics for Personalized Medicine"** are attributed for helpful considerations. 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.