



Research Data Integration of Retrospective Studies for Prediction of Disease Progression

A White Paper

By Erich A. Gombocz

Research Data Integration of Retrospective Studies for Prediction of Disease Progression

Contents

Introduction.....	3
Integration of Multiple Research Data Sources and Medical Observations.....	3
Combining Public Mechanistic Knowledge with Experimental Correlations	5
Finding and Classifying Disease Markers in Context of the Biological System.....	6
Applying a Focused Network Model in Patient Screening to Predict Outcome.....	7
Understanding Biology: Implications for Healthcare and Personalized Medicine	9
Glossary.....	10
References	11

Introduction

While the ability to access, search and meaningfully integrate disparate scientific data and medical information across research and clinical resources alone already is a major challenge for both, the life sciences industry and the medical community at large, it even is a much more demanding effort to do this in context of the relationship of all data to each other within the framework of functional biology. Connecting to different types of databases, querying their contents and integrating the results using semantic methods not only increases the speed and efficiency of data integration and related research by enabling scientists to quickly access and connect relevant information, it also provides the foundation to use retrospective studies for prediction of disease progression on the current patient population.

This white paper provides a high level overview how semantic data integration can be successfully applied to screening and confident decision support in patient care, effectiveness of treatment assessment and for personalized medicine – thereby saving time, costs and lives. References are provided for insight into technical terms (ontologies, semantic mapping, sub-networks, etc.) that are used but not defined here. Specifically, this document outlines the application of IO Informatics' Sentient Suite of software applications ("Sentient") employing semantic methods for dynamic integration of experimental and clinical results, further enriched by semantic integration of public formal mechanistic knowledge networks under a common concept (ontology). Such approach produces interconnected patterns that generate new hypotheses and valuable search models for complex biological functions. Deep relationship-based integration makes it possible to efficiently add new dimensions to research and provide the insights required for sound biomarker-based screening, definition of patient-centric disease stages and applying this knowledge towards prediction of disease progression and optimum treatment options.

Integration of Multiple Research Data Sources and Medical Observations

The Sentient Suite provides secure, HIPAA-compliant web-based access to virtually any data source, including integration of files as well as via direct connections to various databases or laboratory instruments. Data are characterized and classified according to their type and content. The Web Query application allows the user to traverse all those data sets using common identifiers and to set up routine queries to be run instantly. From any set of query results, clinicians and researchers can export the retrieved information to standard tabular formats, but more importantly, informaticians can also open the results directly in the Sentient Knowledge Explorer where data are represented in a graphical network of their interactions and relationships to each other.

Researchers and clinicians are inundated with a huge diversity and quantity of data and information. They commonly need the ability to take self-produced datasets and easily integrate them with other related information of interest, and they need to be able to distill what fits a particular goal. The Sentient Knowledge Explorer allows this by uniting multiple sources in one location, and by mapping diverse data sets to common concepts. At this stage, thesauri are applied to consolidate terminologies (which differ significantly between clinics, biology, and even in-between data sources with similar data of different origin), to deal with synonyms and other "imprecise" data descriptions in order to achieve a high degree of coherence at minimal user curation requirements. This approach is essential when dealing with multiple retrospective studies towards understanding of complex, multifaceted diseases in heterogeneous patient populations.

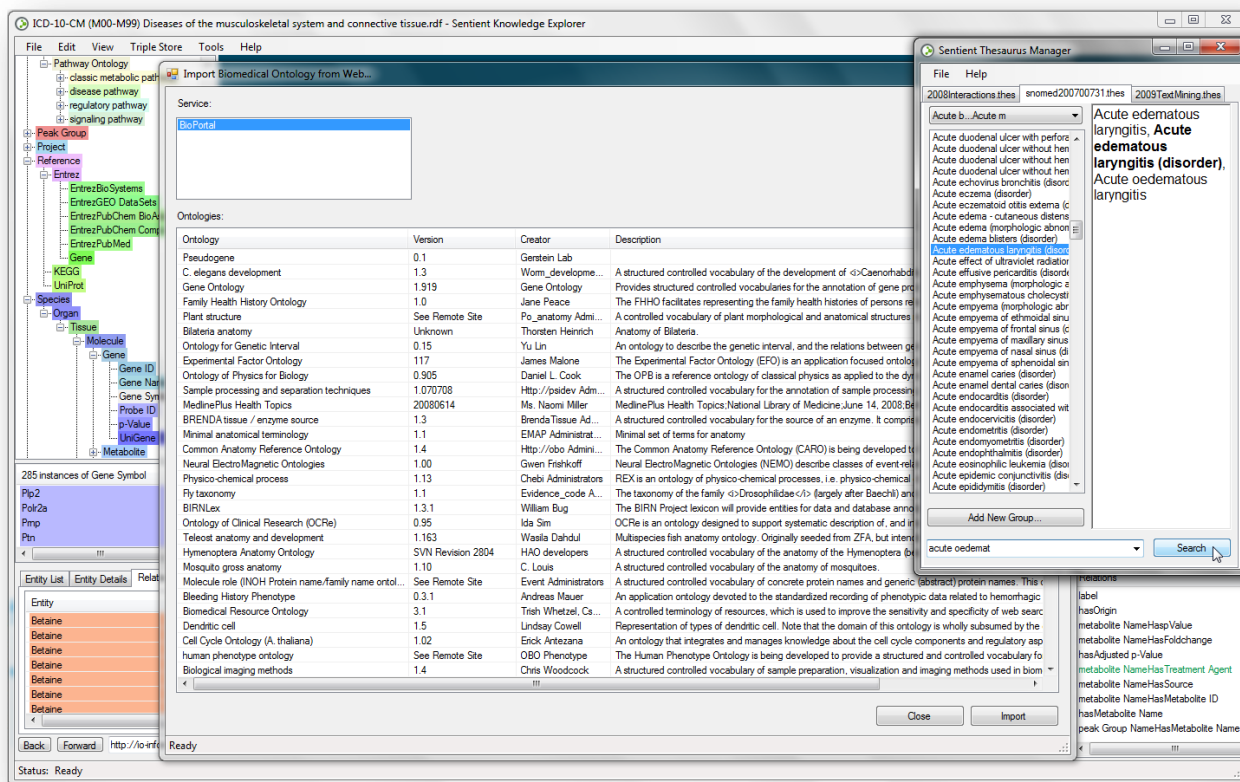


Figure 1: Consolidating terminologies and content: A common classification concept in Sentient Knowledge Explorer, using Thesaurus Manager

In contrast to a traditional relational data warehouse, this technology does not require to *a priori* define which data in which format should be included. Semantic integration is based on Resource Description Framework (RDF), a standard defining data as related to other data. It builds on dynamic, flexible and extensible structures which excel for changing needs, allow reusing and repurposing the data and providing efficient, disambiguous queries. Such considerations are particularly important when dealing with retrospective studies, additional new datasets or changing research objectives. As resources and relationships are stored, navigation through the data presented as interconnected network graphs helps discovering what researchers and clinicians did not even consider searching for. Ambiguities are resolved through the fact that data are looked at in context of other data. Through simply traversing biological concept ontologies visually, hidden relationships become obvious and query by meaning using inference and reasoning becomes possible. The Sentient Suite of software provides the toolset to fully take advantage of these capabilities and applying them successfully to establish disease signatures, validate drug efficacy and toxicity, and stratify patients for clinical trials or prophylactic and therapeutic treatment according to their specific response profiles. Biomarker discovery, qualification and validation from a combination of multiple – OMICS results (such as, for instance, gene expression profiling, protein and metabolic profiling), clinical blood tests, medical imaging endpoints from biopsies, demographic patient data and patient phenotyping is achieved more rapidly and at much lower cost than with alternative solutions.

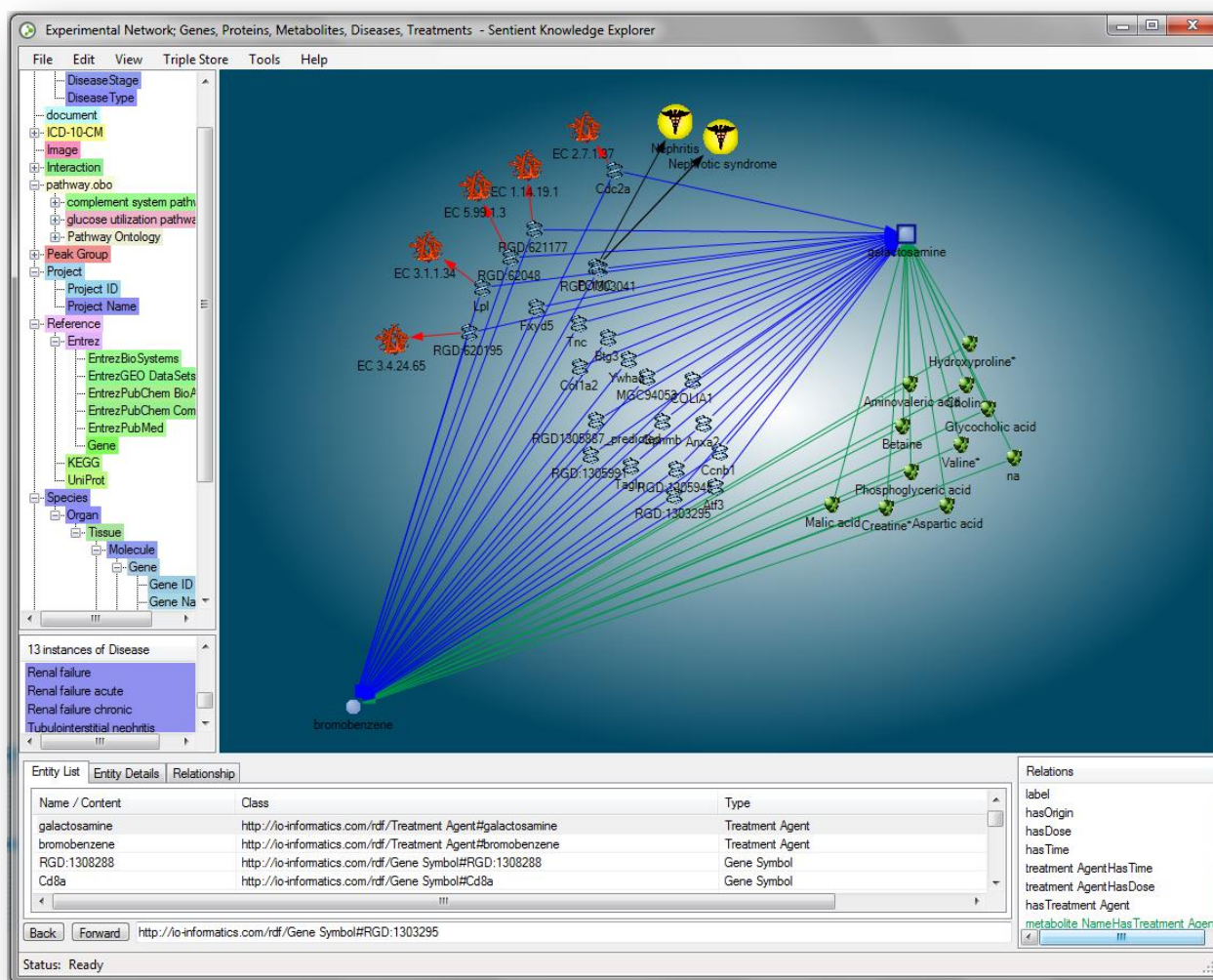


Figure 2: Experimental & clinical network

Figure 2 gives a powerful testimonial for the effectiveness of integration and the clarity of interpretation of the resulting complexity-reduced sub-network. In this example, genes (blue helix icons) and metabolites (green small molecule icons) have been observed to be significantly changed in expression under specific treatments and diseases (yellow medical icons). This helps identifying classifiers and marker panels of experimental relevancy, which need to be confirmed via functional mechanism to ensure that what makes sense from a statistical viewpoint also makes sense from the biological function.

Combining Public Mechanistic Knowledge with Experimental Correlations

Once experimental correlations are established, researchers can begin identifying common pathways, information or processes present in their data. This insights can be gained from connecting and integrating a vast array of publically available, high quality curated data sources such as those from NCBI Entrez with over 30 different scientific sources of biological functions, diseases, literature or chemical structures. Other examples for resources are UniProt, HMDB, KEGG or other pathway-related databases as well as FDA's Clinical Trials database or similar. This can be done effortlessly directly from the Sentient Knowledge Explorer, which connects,

queries and retrieves all relevant information directly via internet access. If desired, formal ontologies, such as those available from NCBO's BioPortal, can be also directly imported and all data mapped to them. The results from those sources are automatically related to the experimental and clinical data network during import, enriching the already rich knowledge further with functional and mechanistic insights from scientific findings reported in reviewed journals. For example, this could lead to the discovery of specific biological pathway involvement which has been uncovered through linking different pathways to enzyme and gene data described in literature as characteristic in influencing the progression of a disease. Similarly, drug interactions and / or adverse effects for patients with certain genetic predisposition can be predicted based on network queries on related genes.

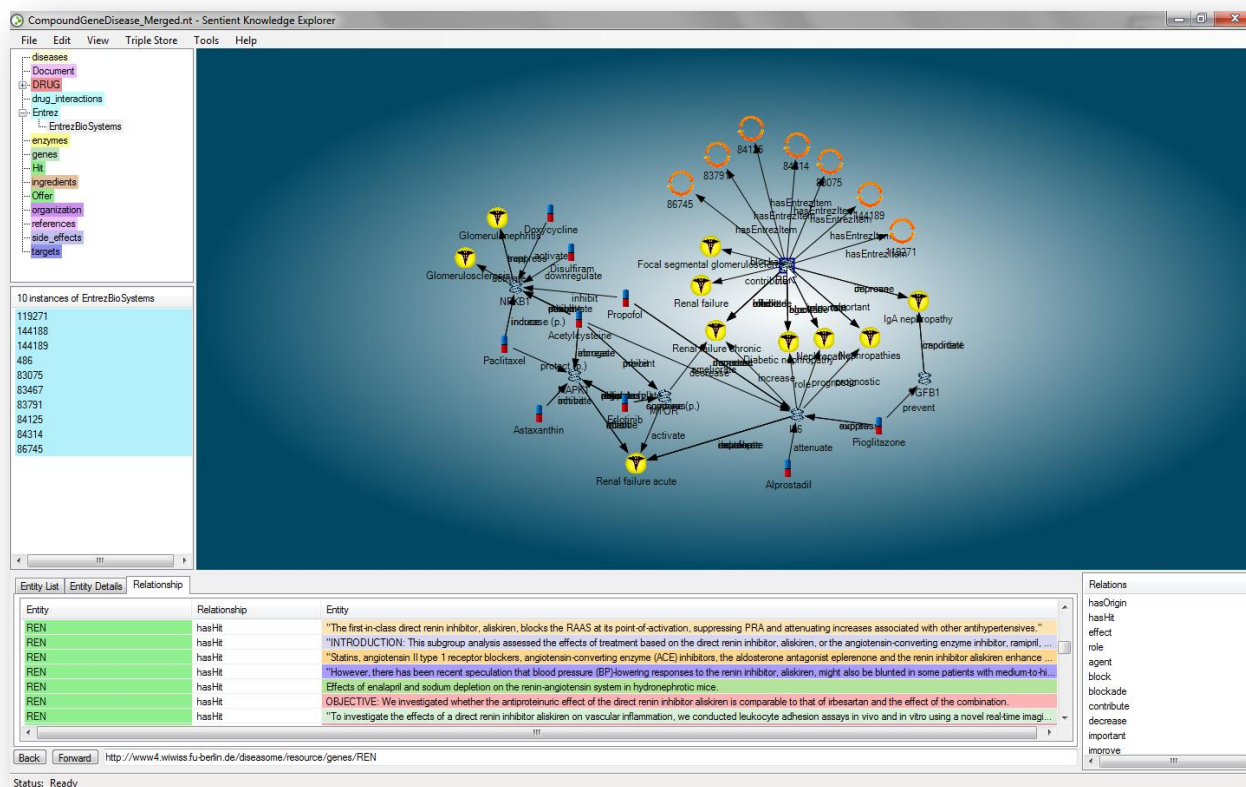


Figure 3: Combining public and experimental knowledge: Biological pathway interaction explanation for observed gene expression changes affecting disease treatment

Finding and Classifying Disease Markers in Context of the Biological System

Once data is loaded into the Knowledge Explorer, researchers and bioinformaticians have the ability to work with consolidated, organized information in one location. The network can be “sliced and diced” according to the questions researchers or clinicians want to ask and only relevant data sub-networks displayed for further exploration. By defining what researchers want to see without knowing all the relationships involved, hypothesis can be easily tested (for instance, inclusion or exclusion of common relationships, setting constraints on numeric values, weighing of relationships according to different classifiers, etc.). Using the graphical SPARQL (SPARQL Protocol and RDF Query Language) query tool in Knowledge Explorer, sophisticated network queries with ranges can be directly generated visually off the graph by selecting sets of nodes. This way, model building and refining is made easy and results can be evaluated immediately.

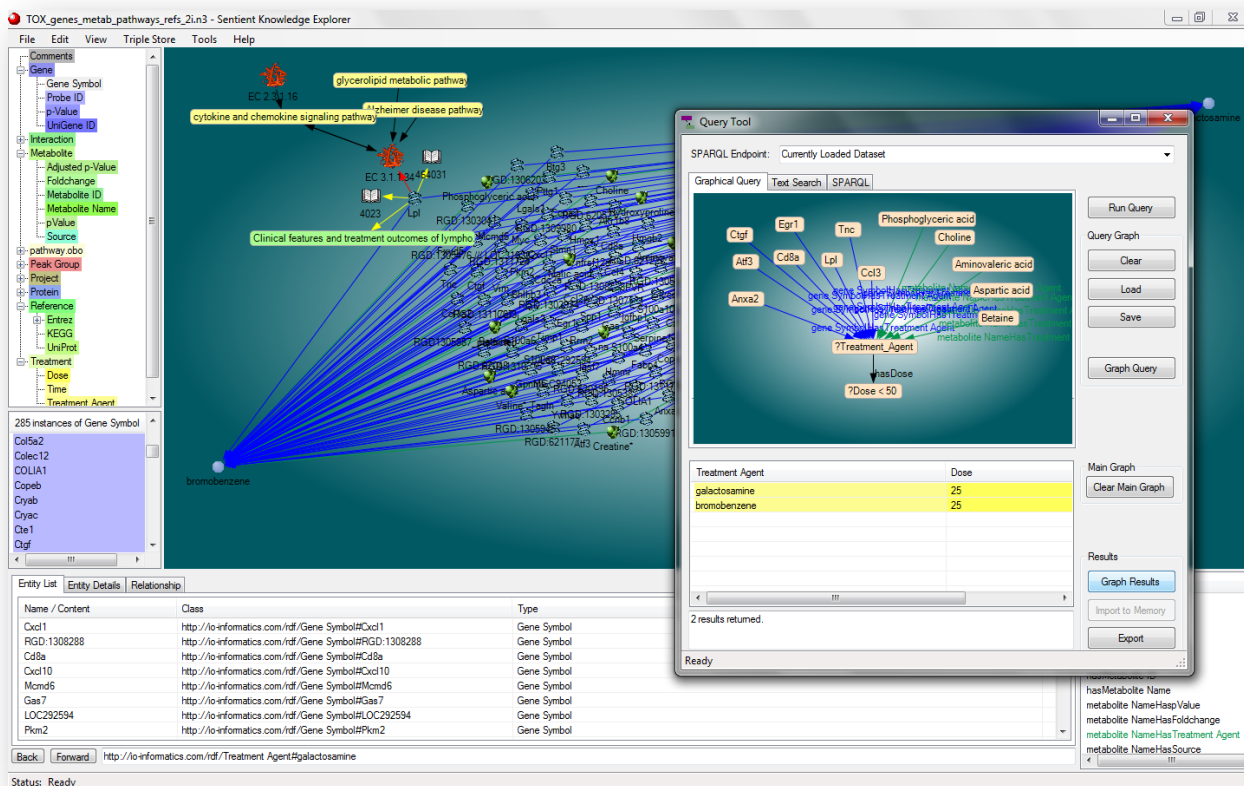


Figure 4: Building systems biology-based biomarker models: Visual generation of toxicity profiles to classify hepatotoxicity

The example in Figure 4 shows how profiles describing specific types of hepatotoxicity have been generated graphically. The pattern in the Query window represents the gene- and metabolite-based biomarker set with their corresponding expression limits; the in the background generated formal SPARQL query can be saved, modified, reloaded and applied to a different or larger dataset without modification.

Applying a Focused Network Model in Patient Screening to Predict Outcome

In the following we will look at examples how arrays of focused network models (each contained in a SPARQL query like the one described above) are used to screen patients for disease states to decide on specific treatment options or to predict disease progression. Collections of network queries provide contained in-depth knowledge which is directly applicable to unknown datasets – thus, they are called “Applied Semantic Knowledgebase” (ASK™).

The screen on the next page depicts such an ASK™ used in treatment selection for prostate cancer patients. As cancer requires aggressive therapy regimens with severe side effects, and treatments are mostly combinations of multiple drugs, prediction of the effectiveness of such treatment can make all the difference in healthcare cost, quality of life for patient or even decides over life and death of the patient. Another application where informed confident decision-making is required is the presymptomatic detection of heart, lung or kidney failures in transplant patients. Organ transplants require a tight immunosuppression therapy with serious side effects and regular biopsies to discover rejection risk at immense costs to the healthcare

system and severe implications for the patient's quality of life. Applying ASK to predict organ failure using a combinatorial blood biomarker test, saves time, money and lives.

Choose Array: Cancer therapies

ASK Arrays

A NHT treatment, early stage downregulation effect [Gene BMs]
 B NHT treatment, time comparison high downregulated [Gene BMs]
 C NHT treatment, timepoint and experiment [32 Gene BMs]
 D NHT treatment, time-independent high upregulated [Gene BMs]
 E Treatment effects, strong up-regulated genes [multi-platform]
 F Treatment effects, strong down-regulated genes [multi-platform]
 G R1881 w/ Lipofectamine, high downregulated [Gene BMs]
 H Treatment comparison, consistently up-regulated genes

[Treatment effects, strong down-regulated genes \[multi-platform\]](#) [Open in Knowledge Explorer\(RQ\)](#)

Your search returned 15 record(s)
 Export: [Excel](#) [Text](#) [TSV](#) [HTML](#) [XML](#) [Open in Knowledge Explorer](#) [\(FQML\)](#) [Chart](#)

ComparisonID	GeneSymbol	FoldChange	Gene_Description	Score
AR_siRNA_RvsNegR	PLA2G2A	-19.5934561994912	Phospholipase A2- membrane associated precursor (EC 3.1.1.4) (Phosphatidylcholine 2-acylhydrolase) (Group IIA phospholipase A2) (GLIC sPLA2) (Non-pancreatic secretory phospholipase A2) (NPS-PLA2). [Source:Uniprot/SWISSPROT;Acc:P14555]	0.4065
LORvsLOE	PLA2G2A	-19.5934561994912	Phospholipase A2- membrane associated precursor (EC 3.1.1.4) (Phosphatidylcholine 2-acylhydrolase) (Group IIA phospholipase A2) (GLIC sPLA2) (Non-pancreatic secretory phospholipase A2) (NPS-PLA2). [Source:Uniprot/SWISSPROT;Acc:P14555]	0.4065
AR_siRNA_EvsNegE	SERPINI1	-13.8921693717385	Neuroserpin precursor (Serpin I1) (Protease inhibitor 12). [Source:Uniprot/SWISSPROT;Acc:Q99574]	6.1078
LORvsLOE	SERPINI1	-13.8921693717385	Neuroserpin precursor (Serpin I1) (Protease inhibitor 12). [Source:Uniprot/SWISSPROT;Acc:Q99574]	6.1078
NHT8to9vsNHT0	C4orf18	-13.3050313075417	TCPD2512. [Source:Uniprot/SPTREMBL;Acc:Q6UWH4]	6.695
NHT5to6vsNHT0	C4orf18	-13.3050313075417	TCPD2512. [Source:Uniprot/SPTREMBL;Acc:Q6UWH4]	6.695
AR_siRNA_EvsNegE	C4orf18	-13.3050313075417	TCPD2512. [Source:Uniprot/SPTREMBL;Acc:Q6UWH4]	6.695
AR_siRNA_RvsNegR	C4orf18	-13.3050313075417	TCPD2512. [Source:Uniprot/SPTREMBL;Acc:Q6UWH4]	6.695
LORvsLOE	C4orf18	-13.3050313075417	TCPD2512. [Source:Uniprot/SPTREMBL;Acc:Q6UWH4]	6.695
AlvsNHT0	MSMB	-12.6863047591418	Beta-microseminoprotein precursor (Prostate secreted seminal plasma protein) (Prostate secretory protein PSP94) (PSP-94) (Seminal plasma beta-inhibin) (Immunoglobulin-binding factor) (IGBF) (PN44). [Source:Uniprot/SWISSPROT;Acc:P08118]	7.3137
NegRvsLOR	MSMB	-12.6863047591418	Beta-microseminoprotein precursor (Prostate secreted seminal plasma protein) (Prostate secretory protein PSP94) (PSP-94) (Seminal plasma beta-inhibin) (Immunoglobulin-binding factor) (IGBF) (PN44). [Source:Uniprot/SWISSPROT;Acc:P08118]	7.3137
AR_siRNA_RvsNegR	MSMB	-12.6863047591418	Beta-microseminoprotein precursor (Prostate secreted seminal plasma protein) (Prostate secretory protein PSP94) (PSP-94) (Seminal plasma beta-inhibin) (Immunoglobulin-binding factor) (IGBF) (PN44). [Source:Uniprot/SWISSPROT;Acc:P08118]	7.3137
AlvsNHT0	MSMB	-12.3925399110126	Beta-microseminoprotein precursor (Prostate secreted seminal plasma protein) (Prostate secretory protein PSP94) (PSP-94) (Seminal plasma beta-inhibin) (Immunoglobulin-binding factor) (IGBF) (PN44). [Source:Uniprot/SWISSPROT;Acc:P08118]	7.6075
NegRvsLOR	MSMB	-12.3925399110126	Beta-microseminoprotein precursor (Prostate secreted seminal plasma protein) (Prostate secretory protein PSP94) (PSP-94) (Seminal plasma beta-inhibin) (Immunoglobulin-binding factor) (IGBF) (PN44). [Source:Uniprot/SWISSPROT;Acc:P08118]	7.6075
AR_siRNA_RvsNegR	MSMB	-12.3925399110126	Beta-microseminoprotein precursor (Prostate secreted seminal plasma protein) (Prostate secretory protein PSP94) (PSP-94) (Seminal plasma beta-inhibin) (Immunoglobulin-binding factor) (IGBF) (PN44). [Source:Uniprot/SWISSPROT;Acc:P08118]	7.6075

Show/Hide SQL

Copyright 2006 - 2009, IO Informatics, Inc. (rev. 2.0.0.550)

Local intranet | Protected Mode: Off

Figure 5: Applied Semantic Knowledgebase (ASK™) for patient screening: Comparative effectiveness of combinatorial cancer treatment with minimal side effects

Understanding Biology: Implications for Healthcare and Personalized Medicine

Integrating research data from retrospective studies, parallel studies, collaborations and other clinical and laboratory-experimental resources provides a complete change in abilities for life sciences and healthcare in general. Query by meaning changes the way we search. Imprecise connections in-between data can be used to infer non-obvious relationships. Linked data collaborations and revival of old studies drive knowledge expansion in the pharmaceutical, life sciences and medical field to a new level, making personalized medicine a reality. IO Informatics' Sentient Suite software solutions have been successfully deployed in industrial settings, in academic Centers of Excellence and research hospitals.



Figure 6: From bench to clinic to personalized care: applications spectrum for prediction and decision support in pharmaceutical industry, life sciences and healthcare

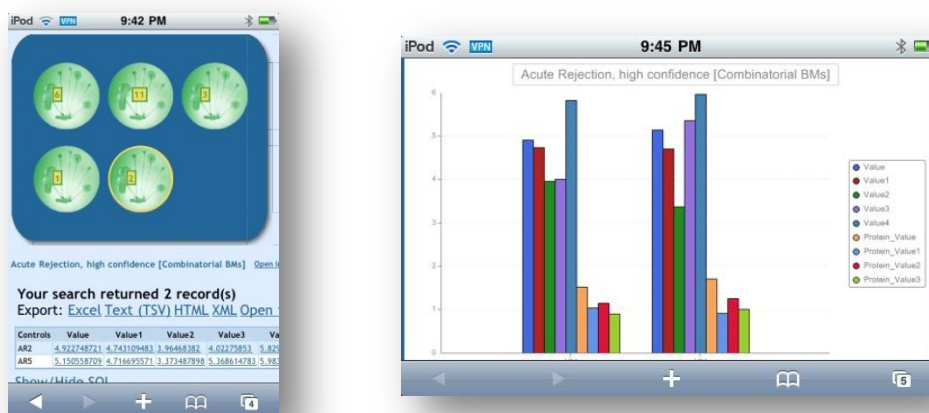


Figure 7: Immediate care: Mobile phone review of patients with high risk for organ rejection

In our world, data are generated at unprecedented speed and in never imagined size and variability. Millions of studies have never been mined or analyzed outside their original purpose, yet they contain a wealth of immensely valuable information when looked at in a different angle. Redoing clinical trials costs a fortune and does not make sense – neither intellectually nor economically.

The described semantic applications make it possible to efficiently leverage data and information across multiple research sources and domain boundaries, utilize retrospective studies in new ways by re-purposing the objectives, and build rapidly dynamic, flexible and extensible solutions which will drive the future of research and discovery.

Glossary

Informatics: The collection, classification, storage, retrieval and dissemination of recorded knowledge.

Ontology: An explicit formal structure of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them.

Resource Description Framework (RDF): Specifications used as a method for the description or modeling of information implemented in web resources.

Relational Database: Database comprising multiple files or related information, usually stored in tables of rows (records) and columns (fields). This allows linking separate files with matching fields so they can be queried simultaneously.

Semantic Network: A graph consisting of nodes that represent physical or conceptual objects and arcs that describe the relationships between the nodes, resulting in a data flow diagram.

Semantic Technology: Tools that encode meaning separately from data or content and allow for inference and reasoning.

References

- (1) E. Gombocz, R. Stanley, J. Eshleman, Z. Rhoades:
"From multiple biomarkers to patient-centric personalized medicine: An 'Applied Semantic Knowledgebase' for decision support"
 Poster at *CHI's Biomarker World Congress 2010*
 Loews, Philadelphia, PA, May 4-6, 2010
 Poster (portrait) [.pdf / 1 MB]
- (2) E. Gombocz, R. Stanley, Z. Rhoades, J. Eshleman:
"Applied Semantic Knowledgebases (ASK ®): Confident decisions in personalized medicine using semantic biology models"
 Poster at *CHI's Bio-IT World Conference 2010*
 World Trade Center, Boston, MA, April 20-22, 2010
 Poster (portrait) [.pdf / 1 MB]
- (3) Z. Rhoades, E. A. Gombocz:
"Charting the Unknown: Capturing and Delivering Value From Understanding Complex Biological Responses"
Am. Biotechnology Laboratory 28 (3): (2010).
- (4) E. A. Gombocz:
"Predictive models for biology in personalized medicine: Are we there yet? "
 Lecture at *Conference on Semantics in Healthcare and Life Sciences (CSHALS)*
 Royal Sonesta Hotel Boston, Cambridge, MA, February 24-26, 2010
 Lecture slides [.pdf / 3.9 MB]
- (5) R. A. Stanley, E. A. Gombocz, Z. Rhoades:
"Realizing personalized medicine with semantic technology: Applied Semantic Knowledgebases (ASK ®) at work"
 Poster at *Conference on Semantics in Healthcare and Life Sciences (CSHALS)*
 Royal Sonesta Hotel Boston, Cambridge, MA, February 24-26, 2010
 Poster (portrait) [.pdf / 1.4 MB]
- (6) E. A. Gombocz, Z. Rhoades:
"Predictive Toxicology: Applied Semantics with major implications towards safer drugs"
 Poster at *SemTech 2009 Semantics Technology Conference*
 The Fairmont Hotel, San Jose, CA, June 14-18, 2009
 Poster (portrait) [.pdf / 1.3 MB]
- (7) E. A. Gombocz, A. J. Higgins, P. Hurban, E. K. Lobenhofer, F. T. Crews, R. A. Stanley, C. Rockey, T. Nishimura:
"Does network analysis of integrated data help understanding how alcohol affects biological functions?" - Results of a semantic approach to biomarker discovery
 Poster at *CHI's Biomarker Discovery Summit 2008* at Loews Philadelphia Hotel, Philadelphia, PA, September 29-October 1, 2008
 Poster (portrait) [.pdf / 1.6 MB]
- (8) E. A. Gombocz, Toshiro Nishimura, Chuck Rockey :
"Towards better understanding of complex biology: Ontology merging across data sources using multiple thesauri in semantic networks"
 Poster at *CHI's Biomarker World Congress 2008* at Loews Philadelphia Hotel, Philadelphia, PA, May 19-21, 2008
 Poster (portrait) [.pdf / 1.0 MB]
- (9) E. A. Gombocz, A. J. Higgins, R. A. Stanley:
"Using semantics in biomarker discovery: Unified cross-OMICS correlation networks help scientists understanding biological functions"
 Poster at *CHI's Molecular Medicine Tri-Conference 2008* at Moscone North Convention Center, San Francisco, CA, March 26-27, 2008
 Poster (portrait) [.pdf / 1.5 MB]