Contextual understanding of experimental data via formal semantic integration of NLP-extracted content with other semantically integrated resources

Erich Gombocz [1)], David Milward[2)], Jason Eshleman [1)]

[1)] IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, USA
[2)] Linguamatics Ltd., St. John's Innovation Centre, Cowley Rd, Cambridge, CB4 0WS, UK

Topic Area: **Biomolecular Semantics**

Summary

Biological systems are inherently complex. Experimental results, especially if they cover multiple experimental modalities or diverse biological responses, are difficult to interpret out of context. This is a key area for the application of semantic technologies. Extensible semantic standards such as RDF, N3 and OWL are used to create coherent, dynamically extensible and re-mappable data models, or "ontologies", for data integration. A common first step is the integration of analytical results under a well-formed ontology. This first step supports the rapid creation of coherent experimental correlation networks and provides a statistically relevant view of system perturbations. However, this does not necessarily provide a better understanding of biological functions involved.

In order to achieve contextual understanding, these networks need to be further enriched by adding mechanistic knowledge. This contextual understanding requires the ability to bring in resources (either through direct connections or via queries to SPARQL endpoints) relevant to biological functions. The addition of information about interactions, pathways or other information from previous observations is relevant to describe biological processes, which may otherwise be missed. Natural Language Processing (NLP) can be used to extract relationships between concepts from resources such as scientific journal articles, collaborations, comparative studies and clinical trials. Using thesauri to harmonize classes and relationships from those extracts and merging them into a dynamically extended ontology provides functionally connected experimental results. When the NLP extracted relationships are semantically integrated with experimental findings, understanding of the biological system is enhanced. This approach makes it possible to derive biomarker patterns or molecular signatures from the network to answer complex biological questions, and also to apply them actively for screening and decision support.

This poster describes a use-case in which multiple experimental datasets (micro-RNA, sequencing, gene expression, drug target assays) have been semantically integrated, enriched with public knowledge resources (tissue-specific gene expression and regulation [TIGER], human RNA drug targets [TargetScan], miRBase, Microcosm, Diseasome) and supplemented with NLP extracted relationships concerning specific diseases (in this case, severe renal failure) from a variety of articles and other sources. Tools used in this scenario were IO Informatics' Sentient Knowledge Explorer for the semantic integration of experimental data, ontology import, network visualization and graphical SPARQL queries in conjunction with relationships extracted from MEDLINE abstracts by Linguamatics I2E enterprise text mining platform.

The resulting semantic network provides a reliable qualification of drug targets with broader applications. The kidney-disease related profiles generated in this example are based on contextual understanding of the biological functions involved in the disease and their manifestation in grounded experimental observations as well as through verification with mined content from trusted resources. Such methodology significantly impacts the way life sciences' research discovers and targets more effective drugs, and is gaining widespread use in personalized medicine.

Introduction & Challenge

- Extraction, translation and loading of heterogeneous data towards a coherent integration still remains a major challenge
- Even then, statistical correlation of experimental findings alone may lead to wrong conclusions when looked at without understanding their mechanistic relevancy to a particular biological response
- Further integration with public knowledge resources almost always requires a harmonization of data classes, unification of ontological terms, thesauri and resource indicators to account for synonyms, and relationship nomenclature or specification differences
- A reliable qualification of experiments for application in biological use cases such as biomarker discovery requires a combination of interconnected mechanistic resources and text-mined reference publications to put the results in a viable functional systems biology context. As such, this correlation of data and resource information provides the understanding for viable decision support

Methodology & Approach

- Using IO Informatics' Sentient Knowledge Explorer, genomic data on liver diseases were semantically integrated with reference data from 5 public resources (tissue-specific gene expression and regulation [TIGER], human RNA drug targets [TargetScan], miRBase, Microcosm, and Diseasome) in a common application ontology knowledge network
- Linguamatics I2E 3.1.1 provided the relationship extraction from MEDLINE abstracts. I2E is an agile NLP-based text mining system that aids in discovery and knowledge synthesis from unstructured text in large document collections. Through the plug-in of large terminologies (Entrez Gene, SNOMED, MeSH, MedDRA, etc.) concepts can be found via their synonyms, and relationships found via co-occurrence within regions of a document, or via precise patterns capturing different linguistic constructions
- Extraction, mapping, and conversion to RDF were accomplished using IO Informatics' Sentient Knowledge Explorer (v.3.3) and the resulting semantic knowledgebase was saved in AllegroGraph (v.4.1.1, Franz, Inc.) semantic datastore

Results & Discussion

- This study focused on indirect relationships between drugs, renal disease, and the genes involved in specific disease treatments. On the NLP-part, this was done by two separate queries: relationships between genes and renal disease and relationships between compounds and those genes
- 'Gene to Disease' used a library query to look for genes (in Entrez Gene) found in a relationship to the MedDRA term "Renal and urinary disorder" looking for multiple syntactic structures including active verbal constructions, passives, nominals and relative clauses.
- 'Compound to Gene' used the same library query, but looked for a "Pharmacologic Substance" from the NCI Thesaurus in a linguistic relationship to the list of genes found by the first query. Fig.1 shows a small section of the output, with the preferred term and concept identifier for the compound, a standardized relationship, and the preferred term and concept identifier for the protein/gene. Both sets of results were exported in tabulated (TSV) format
- Import and mapping of this output into a semantic framework for further enrichment with drug targets, tissue-specific gene expression and regulation to better understand mechanistic aspects of drug interactions on a systems biology-level was performed using IO Informatics' Sentient Knowledge Explorer (KE). The resulting networks and their visual query for drug effects on the disease (directly obtained by node selection in the KE graph without any programmatic steps) are shown in Fig.2.
- Using SPARQL endpoints to 5 linked open data (LOD) sources, the knowledge network in KE was enriched through ontology class, instance and relationship mapping and merging, and stored in AllegroGraph (Franz, Inc.) containing over 3.8 million triples (Fig.3).
- The combined knowledgebase can be directly applied to search for drug targets for renal diseases via visual SPARQL query, providing insights into drug effect on kidney diseases and the genes involved in the treatment response (Fig.4).

Figures

Adalimumab	C65216	block	TNF	7124	1	19949420	1	We also evaluated the potential physiological and therapeutic consequences of TNF-alpha blocking by the biological agent adalimumab, the first fully human (100% human peptide sequences) therapeutic anti-TNF-alpha antibody, on post-translational regulation of TACE.
Adalimumab	C65216	inhibit	TNF	7124	1	19816754	1	We report that the inhibition of TNF-alpha using the neutralizing monoclonal antibody adalimumab has the potential to induce rapid, complete, and long-lasting remission in a life-threatening manifestation of BD.
Alpha-Lipoic Acid	C61595	inhibit	TNF	7124	1	19109223	4	Alpha-lipoic acid inhibits tumor necrosis factor-induced remodeling and weakening of human fetal membranes.
Alpha-Lipoic Acid	C61595	suppress	TNF	7124	1	19563446	1	Further, choline and/or alpha-lipoic acid treatment suppressed TNF-alpha level significantly as compared with that of ovalbumin-challenged mice.

Fig. 1: NLP via Linguamatics I2E text mining: Processing MEDLINE articles

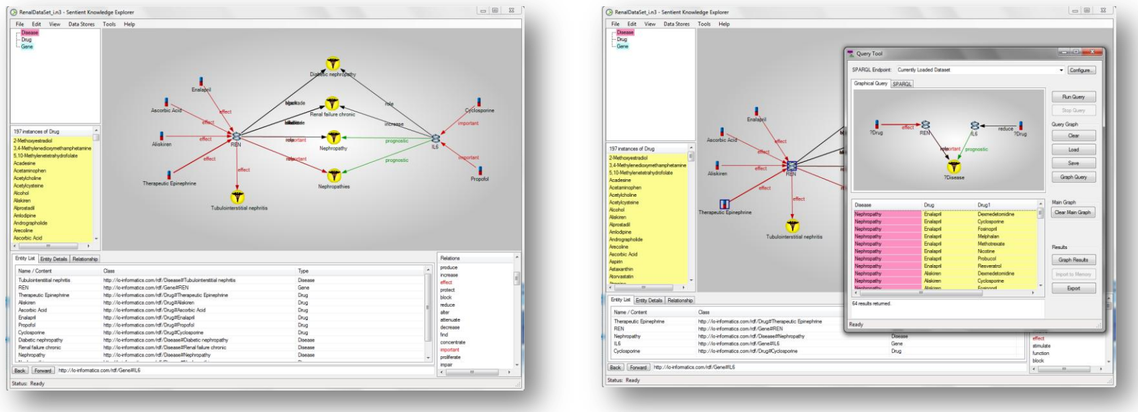

Fig. 2: Kidney disease and drug network from MEDLINE NLP (left) and Visual SPARQL query on drug effects on disease

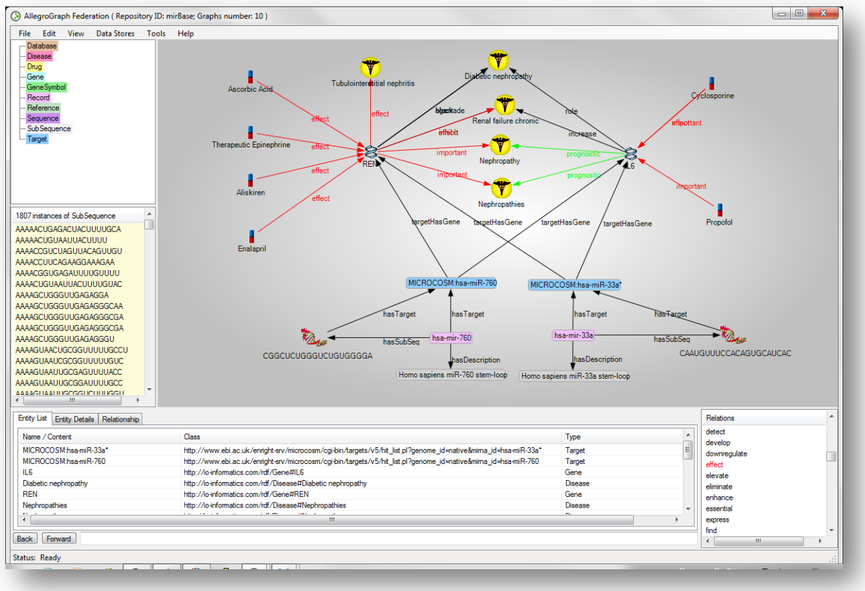

Fig. 3: The merged network: Applicable knowledge from >3.8M Triples, 10 graphs, 5 public resources

A merged network view in Knowledge Explorer:

- Yellow icons represent different renal diseases and genes prognostic or important for those diseases; the blue helix icon are genes involved in the drug response and the pill-style icons represent different drugs affecting the displayed genes.
- The connection to Microcosm (blue) and MiRBase (pink) provide details on the human RNA target sequence

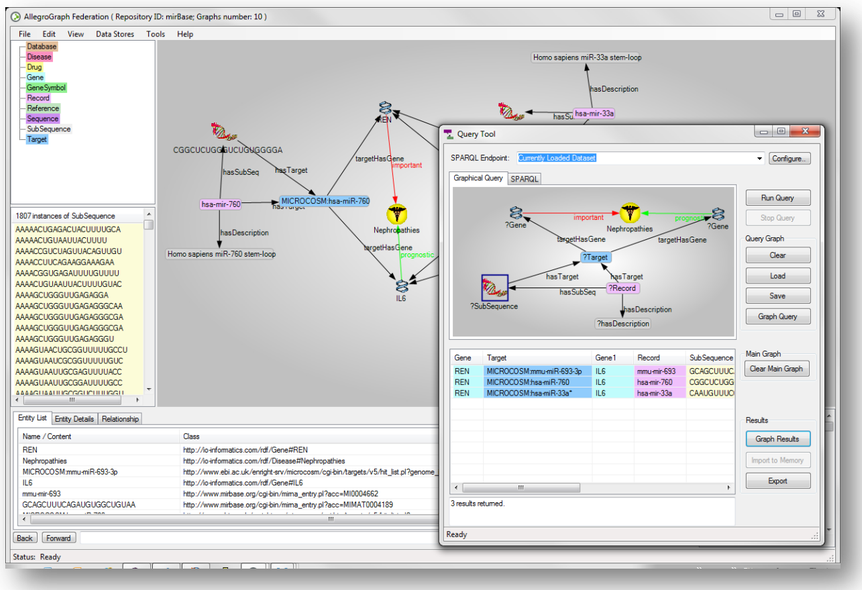

Fig.4: Applying contextualized functional knowledge from all resources:

Payoff from integrated knowledge from multiple resources:

- Visual interrogation allows for model building, refinement and qualification of drug targets in context of their biological responses.
- The resulting knowledgebase can be directly applied as decision support on common and specific drug targets for a broad variety of renal diseases.

Impact

- Applying semantic technology to the integration of experimental and public knowledge resources provides a rapid, cost-effective, extensible and biologically sound method towards understanding and validation of biological functions
- The resulting Applied Semantic Knowledgebases can be directly used for confident decision support in life science, drug discovery and personalized medicine where understanding of the involved biological system is critical.

References

(1) E. Gombocz, R. Stanley, J. Eshleman: "Computational R&D in Action: Integrating Correlation and Knowledge Networks For Treatment Response Modeling and Decision Support", Poster at Advanced Strategies for Computational Drug R&D, Park Plaza Hotel, Boston, MA, Sept. 28-Oct. 1, 2010.
(2) J. Eshleman, E. Gombocz, R. Stanley: "Personalized Medicine in Action: Screening with Semantic Biology Models from Integrated Data Correlation and Knowledge Networks", Poster at CHI's ADAPT 2010, Hyatt Regency Crystal City, Arlington, VA, Sept. 13-16, 2010.
(3) D. Milward, P. Milligan: "Text Data Mining Using Interactive Information Extraction", BioLINK SIG Workshop, ISMB/ECCB (2007).