Title: Integrating Semantics and Numerics: Case Study on Enhancing Genomic and Disease Data Using Linked Data Technologies

Brief Description: We present a novel technique that combines statistical analysis of gene expression data with linked data enrichment to explore the relationships between genes, proteins, drugs, and diseases.

Abstract:
Bioinformatics was an early adopter of semantic technologies and provided many ontologies and datasets in semantic formats to aid integration of biological and biochemical data. There is a lack of tooling support however to help identify and link experimental bioinformatics data and analyses with relevant semantic knowledge. We present a novel approach that combines statistical analyses with semantic data integration to identify support in the literature for findings and highlight where findings expand or contradict knowledge in biomedical databases. We integrate genetic, proteomic, disease, and drug data from numerous sources, including Bio2RDF, Uniprot, Ensembl, and String-DB, along with gene expression data from the Neural Stem Cell Institute and Rensselaer Polytechnic Institute's "Repurposing Drugs with Semantics" project. We will present our integrative system architecture, demonstrate semantically enriched examples from analysis on real datasets showing how semantic representations aid in analysis and interpretation, and discuss architectural lessons learned at scale.

- Integration and reuse of linked datasets and semantic applications
- Provenance of statistical modeling
- Automated data enrichment using semantic datasets, such as Bio2RDF
- Visualization of semantically-enriched experimental data