Ontology driven Semantic Provenance for Heterogeneous Bionomics Experimental Data

Satya S. Sahoo[1], Michael Raymer[1], Cory Henson[1], Amit Sheth[1], Will York[2]
{sahoo.2, michael.raymer, henson.13,, amit.sheth}@wright.edu, will@ccrc.uga.edu
[1]Kno.e.sis Center, Computer Science and Engineering department, Wright State University, Dayton, OH; [2] CCRC, University of Georgia, Athens, GA

Abstract. Scientific experimental data generated by all the bionomic technologies is characterized by heterogeneity in its representation formats, constituents, and generation processes and, therefore, also in its usage. Using the proteomics domain we demonstrate the important role of provenance information o manage, interpret and analyze experimental data. We present a novel approach that employs an ontology as a knowledge model to automatically create semantic provenance information for high-throughput mass spectrometry (MS) data in the glycoproteomics domain. The Semantic Provenance Annotation of Data in protEomics (SPADE) implementation is based on the ProPreO ontology, a large-process ontology (~500 classes, 40 named relationships with 170 class-level restrictions, and 3.1 million instances) that models the complete experimental protocol for MS-based glycoproteomics data analysis. The semantic provenance information created in SPADE enables biologists to query over the semantic provenance information and retrieve exact data using "train-of-thought" expressive queries in SPARQL query language. We also discuss our current work in extending the ProPreO ontology to support toxicological metabolomics experimentation using Nuclear Magnetic Resonance (NMR) spectroscopy. Our strategic goal is to use Semantic Provenance information by pattern recognition and data mining algorithms for comparative or correlation analysis of Liquid Chromatography MS (LCMS) and NMR spectroscopy experimental data as part of toxicological metabolomics studies.

Keywords: Semantic Provenance, ProPreO ontology, Scientific Workflow Proteomics, Glycomics, Toxicological Metabolomics, Biomarkers

1 Introduction

Tandem mass spectrometry (MS/MS), coupled with liquid chromatography (LC-MS/MS) and sophisticated probability-based search algorithms, is a valuable proteomics research tool [9]. In a typical proteomics experiment, the proteins are extracted from the biological material and digested by enzymes to produce peptides [9]. The peptides are partitioned and analyzed by liquid chromatography interfaced to a tandem mass spectrometer [9] (Figure 1). The challenge in giving researchers unified access to the datasets, generated by the ms data analysis process, lies not just in integrating the final results, but correlating and comparing results across processing phases and with multiple constraints.

Provenance information has long been recognized as critical metadata to verify, validate and interpret scientific data. But, in high-throughput experimental processes, such as the ms data analysis process, the associated provenance information is also large in volume. Hence, based on the notion of "computable provenance" [2], we propose the use of semantic provenance for high-throughput experimental data that will enable software application to "understand" and process provenance information. We define semantic provenance as provenance information that refers to a formally defined knowledge model to captures information about the provenance of *data entities* and the *processes* and *agents* that created them.

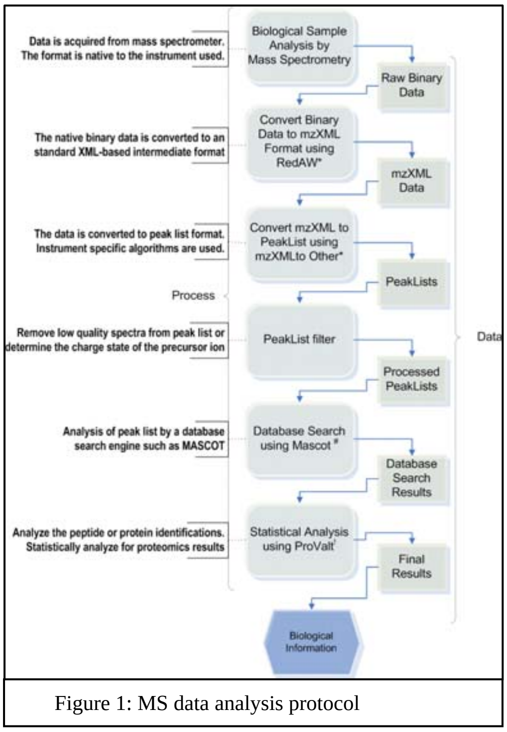
Data is acquired from mass spectrometer. The format is native to the instrument used.
Biological Sample Analysis by Mass Spectrometry
Raw Binary Data
The native binary data is converted to an standard XML-based intermediate format
Convert Binary Data to mzXML Format using RedAW*
mzXML Data
The data is converted to peak list format. Instrument specific algorithms are used.
Convert mzXML to PeakList using mzXMLto Other*
PeakLists
Process
Remove low quality spectra from peak list or determine the charge state of the precursor ion
PeakList filter
Data
Processed PeakLists
Analysis of peak list by a database search engine such as MASCOT
Database Search using Mascot #
Database Search Results
Analyze the peptide or protein identifications. Statistically analyze for proteomics results
Statistical Analysis using ProValt'
Final Results
Biological Information

Figure 1: MS data analysis protocol

Semantic provenance information not only serves to integrate distributed heterogeneous experimental data from multiple phases of high-throughput proteomics protocol, but also enables biologists to pose expressive queries. Those queries explicitly use named relationships, defined in the ProPreO ontology schema, to logically link data entities; hence they closely reflect a biologist's train of thought.

2 ProPreO ontology

The first attempt to formally model the proteomics experimental process was the Pedro UML schema [8]. It soon became clear, as stated in our earlier work [6], that the objectives of the ProPreO ontology are distinct from those of the Pedro UML schema. We therefore engineered ProPreO from the ground up.

Currently, ProPreO models the protein identification process by defining entities corresponding to *datasets*, the *processes* that generate the datasets, and the *agents* that implement the processes (Figure 2). The *named relationships* that interconnect these classes of entities are central [7] to capturing their logical context. ProPreO ontology has been released for public use through the Open Biomedical Ontologies (OBO), a resource of the National Center for Biomedical Ontologies (NCBO).[1]

The SPADE implementation uses a two-phase approach to create semantic provenance based on the ProPreO ontology:

[1] http://www.bioontology.org/ncbo/faces/pages/ontology_list.xhtml.

1. **Extraction of entities**: After identifying "entities of interest" that constitute relevant provenance information, the entities are extracted from the experimental data files and classified as instances of ProPreO ontology concepts. This is implemented at each intermediate step of the workflow, resulting in an aggregated list of ProPreO ontology class instances at the end of the workflow.
2. **Inference of named relationships between the entities**: The named relationship between entities is inferred from the ProPreO ontology schema. We have implemented this reasoning task using the Jena API [5].

The final RDF file populates the provenance model in an Oracle 10g database[*]. We use SPARQL query language for RDF [1] to query the provenance information.

3 Implementation

SPADE is based on the services-oriented architecture (SOA) utilizing semantic Web services (SWS) as components that are composed into a multi-step semantic Web process (i.e., a scientific workflow with SWSs as components, Fig. 3). SPADE is realized using the following components:

1. **The MS data analysis Web process**: Each processing phase of the data analysis protocol is modeled as SWS that are deployed in the Taverna [3] workflow engine.
2. **The semantic provenance modules**: The Semantic provenance modules (SPM) are SWSs that create the provenance information (as described in Section 3.2) and are plugged into the Web process.

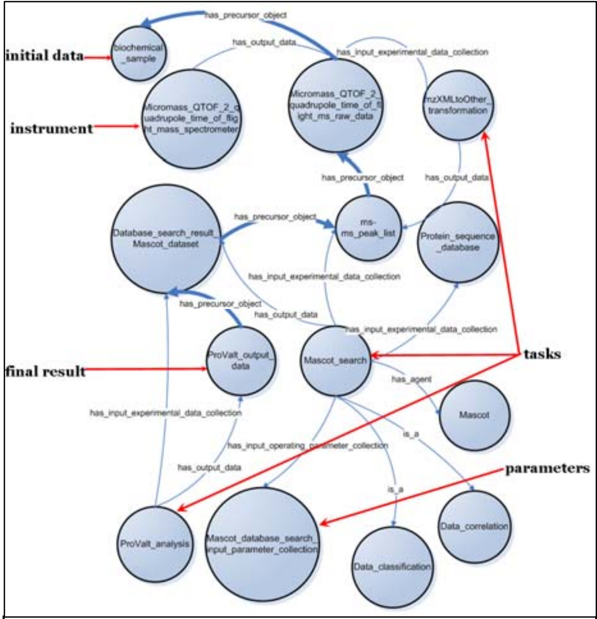
has_precursor_object
has_output_data
has_input_experimental_data_collection
initial data
biochemical_sample
instrument
Micromass_QTOF_2_quadrupole_time_of_flight_mass_spectrometer
Micromass_QTOF_2_quadrupole_time_of_flight_ms_raw_data
mzXMLtoOther_transformation
has_precursor_object
has_output_data
Database_search_result_Mascot_dataset
has_precursor_object
ms-ms_peak_list
Protein_sequence_database
has_input_experimental_data_collection
has_precursor_object
has_output_data
has_input_experimental_data_collection
tasks
final result
ProValt_output_data
Mascot_search
has_agent
Mascot
has_input_experimental_data_collection
has_input_operating_parameter_collection
is_a
has_output_data
parameters
is_a
ProValt_analysis
Mascot_database_search_input_parameter_collection
Data_classification
Data_correlation

Figure 2: ProPreO ontology concepts and relationships

[*] http://www.oracle.com/index.html

4 Results and Discussion

The following categories of queries were executed against the provenance information generated during the sample runs:

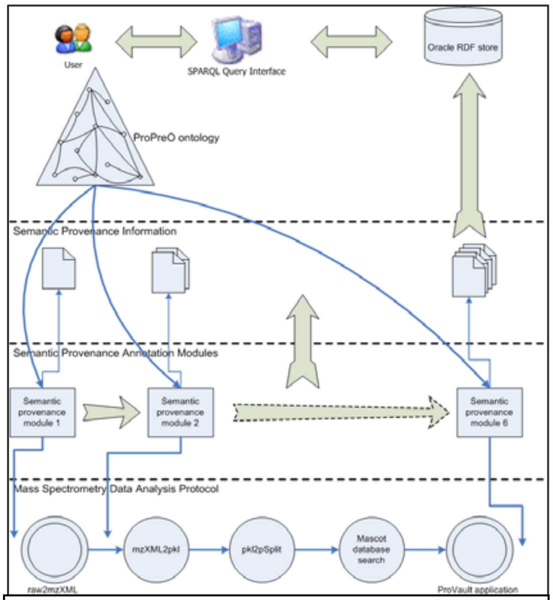
User
SPARQL Query Interface
Oracle RDF store
ProPreO ontology
Semantic Provenance Information
Semantic Provenance Annotation Modules
Semantic provenance module 1
Semantic provenance module 2
Semantic provenance module 6
Mass Spectrometry Data Analysis Protocol
mzXML2pkl
pkl2pSplit
Mascot database search
raw2mzXML
ProVault application

Figure 3: Architecture of the SPADE implementation

1. **Single value constraint:** Find all the RAW files (generated by the ms instrument, Figure 1) that are derived from biochemical samples taken from *T.cruzi* organism.
2. **Multiple value constraint:** Find all RAW files that are generated from biochemical samples derived from organism *Homo sapiens*, have the *profile* format (a parameter for processing of data) and generated by ms instrument with serial number 7635532.
3. **Correlation of datasets using named relationships across multiple intermediary concepts**: Given a specific Provalt (analysis application) result file *JA_Serum_glycopeptides_ALL_txt*, find all its related RAW files. This should identify several files, as Provalt application combines and summarizes data from several different mass spectrometry analysis runs.

We manually cross-checked the query results with the dataset values and found exact match between the query results and dataset values.

5 Toxicological Metabolomics

We propose to extend the ProPreO ontology with concepts and relations related to toxicological metabolomics experimentation using NMR spectroscopy approach [4]. The toxicological metabolomics project aims to identify biomarkers for detection of toxic agents. This involves the administration of toxins to set of rats to study the effect of toxins through the analysis of blood and urine samples. The project involves two parallel analysis methods, one using LCMS approach and the other using NMR spectroscopy approach. The two efforts are synchronized in terms of following similar experimental protocols, method of administering toxin sample to the specimen and collection of tissue samples at same time point.

The two analysis techniques have certain inherent advantages and disadvantages such as:

a) Preparation time: For NMR spectroscopy based analysis, the preparation time is minimal whereas LCMS requires large preparation time

b) Sensitivity: LCMS has better sensitivity as compared to NMR spectroscopy analysis technique

c) Specimen survivability: LCMS approach requires the termination of the specimen whereas NMR spectroscopy approach does not have this requirement

Thus, the comparison of experimental data from each of the two approaches will enable scientists to compare and correlate results to gain vital insights in the toxicological metabolomics study. For example, the absence of an entity in a sample may be due to sensitivity factors, experimental material, behavior of specimen unrelated to toxin metabolic effects (specimen may exhibit self-destructive behavior such as not feeding after administration of the toxin which may not be direct effect of the toxin) that may affect metabolic activity and skew biomarker readings.

Hence, through use of Semantic Provenance information, pattern recognition or mining algorithms can more effectively analyze comparable or related experimental datasets. Semantic Provenance will also enable the scientists to identify the cause of observed effects by tracing the lineage or history of a dataset. To achieve this, we propose to extend ProPreO ontology by incorporating NMR spectroscopy protocol specific concepts and relationships [4]. This will enable the semantic annotation of not only LCMS experimental data but also NMR spectroscopy experimental data. This integrated semantic provenance platform for both LCMS and NMR spectroscopy experimental data will enable scientists to leverage the available data to gain critical research insights in toxicological metabolomics.

6 Conclusion

The research and application presented here lead us to two key observations:

Use of ontology-based provenance information enables knowledge-driven access to distributed heterogeneous experimental data. We demonstrated the central role of ontology and use of semantic provenance information for effective querying.

Use of service-oriented architecture (SOA) based implementation for automated provenance creation for high-throughput experimental processes. We noted the importance of SWS based approach to automatically create semantic provenance creation.

We also discussed our proposed extension of the ProPreO provenance ontology with NMR concepts and relationships to enable an integrated semantic provenance platform to compare LCMS and NMR spectroscopy experimental datasets from toxicological metabolomics study [4].

References

[1] *SPARQL Query Language for RDF*, in A. S. Eric Prud'hommeaux, ed., *W3C Working Draft*, W3C, 2006.

[2] C. Goble, *Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics*, *Workshop on Data Derivation and Provenance*, Chicago, 2002.

[3] D. Hull, Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., and Oinn, T., *Taverna: a tool for building and running workflows of services*, Nucleic Acids Research (2006), pp. W729-W732.

[4] B. K. Kelly, Anderson, P. E., Reo, N. V., DelRaso, N. J. , Doom, T. E., Raymer, M. L., *A proposed statistical protocol for the analysis of metabolic toxicological data derived from NMR spectroscopy*, *7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2007)*, Cambridge - Boston, Massachusetts, USA, 2007, pp. 1414-1418.

[5] B. McBride, *Jena: A Semantic Web Toolkit.*, IEEE Internet Computing, 6 (2002), pp. 55-59.

[6] S. S. Sahoo, Thomas, C., Sheth, A., York, W. S., and Tartir, S., *Knowledge modeling and its application in life sciences: a tale of two ontologies*, *Proceedings of the 15th international Conference on World Wide Web WWW '06* Edinburgh, Scotland, 2006, pp. 317-326.

[7] A. Sheth, Arpinar, I.B., Kashyap V., *Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships*, in B. A. Masoud Nikravesh, Ronal Yager and Lotfi A. Zadeh, ed., *Enhancing the Power of the Internet: Studies in Fuzziness and Soft Computing*, Springer-Verlag, 2003, pp. 63–94.

[8] P. N. Taylor CF, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard , Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates JR 3rd, Brass A, Brown AJ, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG, *A systematic approach to modeling, capturing, and disseminating proteomics experimental data*, Nat. Biotechnol., 21 (2003 Mar), pp. 247-54.

[9] B. Weatherly, Atwood, J., Minning, T., Cavola, C., Tarleton, R., Orlando, R., *A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results*, Mol. Cell. Proteomics, 4 (2005), pp. 762-772.