The BioMANTA Ontology: Integration of Protein-Protein Interaction Data

Andrew Newman[1], Jane Hunter[1], Yuan-Fang Li[1], Chris Bouton[2], and Melissa Davis[3]

[1] School of Information Technology and Electrical Engineering,
[3] Institute for Molecular Biosciences and ARC Center of Excellence in Bioinformatics,
The University of Queensland, Queensland, 4072 Australia
[2] Computational Sciences Center of Emphasis, Pfizer, Cambridge, USA

[1] {anewman,jane,liyf}@itee.uq.edu.au
[2] christopher.m.bouton@pfizer.com
[3] m.davis@imb.uq.edu.au

Abstract. Protein-protein interaction (PPI) and biomolecular pathway data hold tremendous potential for drug discovery and development. However, relevant data sources are currently distributed across a wide range of disparate, large-scale, publicly-available databases, web sites and repositories and are described using a wide range of taxonomies and ontologies. Sophisticated integration, manipulation, processing and analysis of these data sets are required in order to reveal previously undiscovered interactions and pathways that will lead to the discovery of new drugs. The Semantic Web has been investigated as a solution to this problem by a number of projects that use RDF and OWL to integrate, represent and analyze protein interaction data. However, existing applications have suffered from certain limitations that hinder their usefulness. In this paper, we describe work being undertaken within the BioMANTA project that aims to identify and overcome the limitations associated with the application of Semantic Web technologies to protein interaction network analysis. In particular we describe the BioMANTA OWL ontology that has been designed to enable multiple data sources to be integrated within a single RDF triple store through a common PPI model. The primary aim of the BioMANTA ontology is to provide a practical means of facilitating the integration of semantically disparate data sets – it does not aim to provide a precise biological, chemical or physical model of how proteins interact. We also describe how this ontology was developed through the refinement, harmonization and extension of existing ontologies. Finally, we describe the mapping, integration and querying of a range of protein-protein interaction data sets, based on the ontology.

1. Introduction

BioMANTA is a collaborative systems biology project focused on *in silico* drug discovery through integrated data set analysis of molecular interactions and biochemical pathways. This requires real-time analysis and feedback, across large disparate data sets in order to identify lead candidates or new interaction networks. It draws on a large number of publicly available databases of varying size, specialization, coverage and reliability. The data sets of relevance are in different formats, are generated through widely

varying methodologies and are described using very different vocabularies and terminologies.

The Semantic Web technologies, OWL and RDF, were chosen as the platform for representing, modeling and analyzing protein-protein interaction (PPI) and molecular pathway data as they provide a formal and extensible semantics, ideal for data integration [Belleau et al. 2007]. A number of projects including the Gene Ontology [Ashburner et al. 2000] and the OBO ontology [Smith et al. 2007] have previously used RDF and OWL to mark up biological knowledge and enable additional reasoning. In addition, a number of ontologies covering different aspects of systems biology and protein interaction networks have also been developed: biological processes and molecular function (GO), phylogenetic classifications (NCBI taxonomy), sequence annotation (Sequence Ontology), molecular interactions (PSI-MI), cell types (Cell Type Ontology) and pathways (BioPax). However, there does not exist a unified representation of all of the concepts that a biologist might consider relevant to the study of protein-protein interactions within a single OWL ontology.

Within the BioMANTA project, a high-level, extensible ontology has been created to integrate concepts from relevant existing ontologies. This ontology serves as a basis on which knowledge relevant to PPI can be harmonized, providing coverage over a wide conceptual space. It is also used to formalize and validate the properties used in our integration process - a challenging task which demands scalability, efficiency and accuracy over the very large scale data sets involved.

2. The BioMANTA Ontology

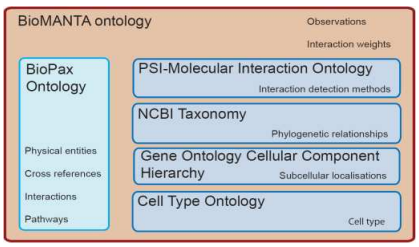

Fig. 1. Overview of BioMANTA Ontology

The BioMANTA ontology [M. Davis et al. 2008][1] is an OWL DL ontology focusing on integrated concepts from: PSI-MI[2], BioPAX level 2 [BioPAX Workgroup 2004], Cell Type [Bard et al. 2005], Gene Ontology [Bard 2005] and NCBI Taxonomy[3]. The ontology combines the use of top-down and bottom-up development - incorporating terms as appropriate and required in order to leverage existing data sets. We have reused vocabularies where appropriate. For example, we incorporated the Cell Type ontology (a

[1] http://biomanta.sourceforge.net/2007/07/biomanta_extension_02.owl

[2] http://psidev.sourceforge.net/mi/rel2/doc/

[3] http://www.ncbi.nlm.nih.gov/Taxonomy/

structured controlled vocabulary[4] taken from OBO Foundry) after it was converted to OWL format[5].

Our approach to knowledge representation combines the three different levels of attitude to data modeling [Ruttenberg et al. 2006]. Our ontology allows us to express: "there exists a protein" (the domain level) and "database A says it has these properties and database B says it has these properties" (the record level). We also align the experimental provenance information (statement level) to improve query quality – allowing us to filter out various experimental types. This work is based on object identification or record linkage [Batini and Scannapieca 2006] which seeks to integrate various data sets across databases and the Semantic Web [Guha 2004].

In the ontology, the hierarchical but expressive properties of PSI-MI are combined with the extensibility and richer relationships available in BioPAX (expressed in OWL). Among others, the ontology consists of the following key concepts:

- Different types of Observation including: Experimental, Predicted, and Inferred; and
- Provenance information including: data source, the type of experiment, the cell type, inferencing method, sub-cellular location and observation reference (a BioPAX publication cross reference).

While BioPAX is expressed in OWL there are numerous issues with its modeling approach [Ruttenberg et al. 2006] [Ruttenberg et al. 2005] [Motik et al. 2007]. One of the most detrimental aspects, with respect to our requirements, was the lack of context or meaning when the `openControlledVocabulary` class is used to link to external terms. This was overcome by using URIs instead of string literals. We also developed a process which involved converting (the Cell Type ontology) from OBO to OWL and defining classes and instances to represent these richer relationships.

3. The Integration of Biological Data Sets

Within the BioMANTA project, various PPI-related data sets are integrated to form a uniform RDF representation that enables answering of complex queries. In this section, we present details of the integration process.

In systems biology, a number of protein databases such as UniProt [Wu et al. 2006], DIP [Salwinski et al. 2004], IntAct [Kerrien et al. 2007] and MPact [Guldener et al. 2006] have been developed. They contain partially overlapping information about a wide variety of proteins/genes. However, most of these data sets employ their own naming conventions. For example, the protein identified as "27628" in DIP is the same protein as the one identified as "115 dax human" in IntAct. It is important to be able to refer to the real protein by either identifier and to be able to query and retrieve its properties from both data sets. Within the various data sets there is also duplication, inconsistency and noise. For example, we have noticed that in one particular data set, multiple different proteins have been mapped to the same ID. Hence, relying on the matching of names from different data sets is not entirely reliable. For this reason, we decided to integrate the databases using a combination of the UniProt ID and genomic sequence to uniquely identify a protein. These two values are then associated with a protein in-

[4] http://obofoundry.org/cgi-bin/detail.cgi?id=cell

[5] http://biomanta.sourceforge.net/2007/07/celltype_instance_edit.owl

stance, without a name from any particular database but instead, using a blank node. The integration process can be conceptually viewed as the following set of steps.

1. **PSI-MI to RDF translation -** XML data sets in PSI-MI [Hermjakob et al. 2004] format are translated to RDF. In this step, all the information associated with proteins and interactions in PSI-MI is modeled using RDF constructs, concepts and properties defined in the BioMANTA ontology. For example, the organism, the local identifiers and the genomic sequence information are captured as RDF triples.
2. **UniProt ID augmentation** - UniProt is a comprehensive protein database. We decided to use UniProt IDs in the integration process to merge proteins from different data sets. However, not all data sets contain UniProt IDs. In this step, with the help of external mapping files between local IDs and UniProt IDs, a UniProt ID is added to each protein instance.
3. **Sequence augmentation –** Proteins that have genomic sequences can be disambiguated by evaluating the sequence, using such tools as Blast, and equivalent sequences are used to identify proteins across data sources. In this step, all the missing sequences of proteins are added to the RDF instances from external mapping files for individual data sets.
4. **Protein integration** - Matching UniProt IDs and sequences, the final step merges proteins from different data sets into a single, uniform representation. Proteins with different UniProt IDs are considered to be different ones; those with same UniProt IDs but different sequences generate warnings; proteins with matching IDs and sequences will be merged into a single protein, together with their annotations.

Figure 2 depicts a merged protein with UniProt ID "Q12522", it shows the result of merging instance data from the MIPS and UniProt databases. We have kept the various synonyms from both databases: "CDC95" and "elF6". This allows queries that only use one of the databases to find data related to the protein, in another database. Note that ovals represent RDF blank nodes and squares represent RDF URI references or literals.

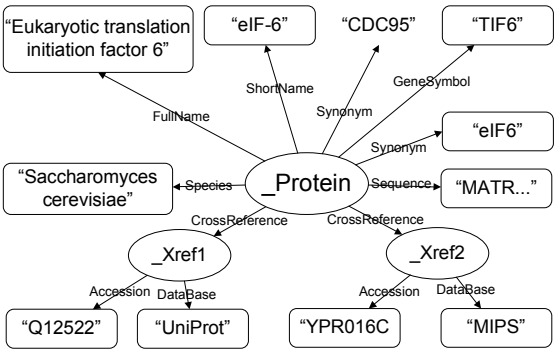

Fig. 2. A protein instance, merged from UniProt and MIPS, using UniProt ID and sequence.

As the BioMANTA ontology uses concepts from NCBI the integrated protein data can be used for inferring certain relationships such as "all mammalian interactions" as well as determining the quality of the observation based on experiment type. This reasoning can be employed to filter the data and improve its quality during the integration process by only incorporating observations that have occurred within mammals.

Although the aim of the BioMANTA project is to enable drug discovery and to identify therapeutic targets, data sets about yeast (Saccharomyces cerevisiae, NCBI Taxonomy ID 4932) have also been used to evaluate the feasibility of the approach - because of their relative simplicity.

4. Conclusions

The BioMANTA project aims to apply Semantic Web technologies to the modeling, integration, analysis and querying of protein-protein interaction data. An OWL ontology has been developed to overcome limitations of existing ontologies, and to better enable the representation, integration, querying of and reasoning across complex protein-protein interactions and molecular pathways. Based on the BioMANTA ontology, an integration process has been designed and implemented to translate, augment and integrate different complementary data sets on protein-protein interactions and biological pathways, in order to assist biologists with the discovery and design of new drugs.

References

[Ashburner, M., et al. (2000)] "Gene Ontology: tool for the unification of biology." Nature Genetics **25**: 25-29.

[Bard, J., et al. (2005)] "An ontology for cell types." Genome Biology **6**(2).

[Bard, J. a. R., S.Y. and Ashburner, M. (2005)] "An ontology for cell types." Genome Biology **6**: R21.

[Batini, C. and M. Scannapieca (2006)] Data Quality. Berlin, Germany, Springer-Verlag.

[Belleau, F., et al. (2007)] Bio2RDF: Towards A Mashup To Build Bioinformatics Knowledge System. Health Care and Life Sciences Data Integration for the Semantic Web, 16th International World Wide Conference, Banff, Alberta, Canada.

[BioPAX Workgroup (2004)] BioPAX – Biological Pathways Exchange Language (Level 2, Version 0.5 (Draft Release) Documentation), BioPAX Workgroup.

[Guha, R. (2004)] Object co-identification on the Semantic Web. 13th World Wide Web Conference, New York, USA.

[Guldener, U., et al. (2006)] "MPact: the MIPS protein interaction resource on yeast." Nucleic Acids Res **34**(Database issue): D436-41.

[Hermjakob, H., et al. (2004)] "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data." Nat Biotechnol **22**(2): 177-83.

[Kerrien, S., et al. (2007)] "IntAct--open source resource for molecular interaction data." Nucleic Acids Res **35**(Database issue): D561-5.

[M. Davis, et al. (2008)] Integrating Hierarchical Controlled Vocabularies with OWL Ontology: A Case Study from the Domain of Molecular Interactions. 6th Asia Pacific Bioinformatics Conference (APBC08), Kyoto, Japan.

[Motik, B., et al. (2007)] Bridging the gap between OWL and relational databases. Proc. of the 16th International World Wide Web Conference, Alberta, Canada, ACM Press New York, NY, USA.

[Ruttenberg, A., et al. (2005)] Experience Using OWL DL for the Exchange of Biological Pathway Information. OWL: Experiences and Directions Workshop, Galway, Ireland.

[Ruttenberg, A., et al. (2006)] What BioPAX communicates and how to extend OWL to help it. OWL: Experiences and Directions Workshop Series, Athens, Georgia, USA.

[Salwinski, L., et al. (2004)] "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-51.

[Smith, B., et al. (2007)] "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nature Biotechnology **25**: 1251-1255.

[Wu, C. H., et al. (2006)] "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-91.