Protein Ontology Development using OWL

Amandeep S. Sidhu[1], Tharam S. Dillon[1], Elizabeth Chang[2], and Baldev S. Sidhu[3]

[1]Faculty of Information Technology, University of Technology, Sydney, Australia
{asidhu, tharam}@it.uts.edu.au

[2]School of Information Systems, Curtin University of Technical University, Perth, Australia
Elizabeth.Chang@cbs.curtin.edu.au

[3]State Council of Education Research and Training, Punjab, India
bsidhu@biomap.org

Abstract. To efficiently represent the protein annotation framework and to integrate all the existing data representations into a standardized protein data specification for the bioinformatics community, the protein ontology need to be represented in a format that not enforce semantic constraints on protein data, but can also facilitate reasoning tasks on protein data using semantic query algebra. This motivates the representation of Protein Ontology (PO) Model in Web Ontology Language (OWL). In this paper we briefly discuss the usage of OWL in achieving the objectives of Protein Ontology Project. We provide a brief overview of Protein Ontology (PO) to start with. In the later sections discuss why OWL was an ideal choice for PO Development.

Keywords: Protein Ontology, Biomedical Ontologies, OWL based Protein Ontology, Protégé, OWL, Proteomics, Data Integration.

1. Background

Traditional approaches to integrate protein data generally involved keyword searches, which immediately excludes unannotated or poorly annotated data. It also excludes proteins annotated with synonyms unknown to the user. Of the protein data that is retrieved in this manner, some biological resources do not record information about the data source, so there is no evidence of the annotation. An alternative protein annotation approach is to rely on sequence identity, or structural similarity, or functional identification. The success of this method is dependent on the family the protein belongs to. Some proteins have high degree of sequence identity, or structural similarity, or similarity in functions that are unique to members of that family alone. Consequently, this approach can't be generalized to integrate the protein data. Clearly, these traditional approaches have limitations in capturing and integrating data for Protein Annotation. For these reasons, we have adopted an alternative method that does not rely on keywords or similarity metrics, but instead uses ontology. Briefly, Ontology is a means of formalizing knowledge; at the minimum ontology must include concepts or terms relevant to the domain, definitions of concepts, and defined

relationships between the concepts. Ontology for Protein Domain must contain terms or concepts relevant to protein synthesis, describing Protein Sequence, Structure and Function and relationships between them. Protein Ontology (PO) provides clear and unambiguous definitions of all major biological concepts of protein synthesis process and relationship between them using OWL. The use OWL in PO provides a unified controlled vocabulary both for annotation data types and for annotation data. We have built PO **[Sidhu et al., 2006, Sidhu et al., 2005a, Sidhu et al., 2005b, Sidhu et al., 2005c, Sidhu et al., 2004a, Sidhu et al., 2004b, and Sidhu et al., 2004c]** to integrate protein data formats and provide a structured and unified vocabulary to represent protein synthesis concepts. PO also helps to codify proteomics data for analysis by researchers. The Complete Class Hierarchy of Protein Ontology (PO) is shown in **Figure 1.** More detailed UML Diagrams for PO are available at the website: **http://www.proteinontology.info/**

A XML Database of 10 Major Prion Proteins available in various Protein data sources, based on the vocabulary provided by Protein Ontology is available on the PO website. Soon we will have all the 57 Prion Proteins known to exist, and user interfaces to browse and query the database. The XML database currently contains 24 tables, 261 attributes and 17550 instances. Prion Protein is a membrane bound protein of 253 amino acid residues in length that is normally found in neurons and several other cell types. The abnormal Prion Protein is resistant to digestion with enzymes that breaks down normal proteins, and accumulates in the brain. Abnormal Prion Proteins are the major cause of various Human Prion Diseases in Brain like Fatal Familial Insomnia. Recently, discovery of Interesting Properties of Prion Proteins encouraged Scientists to understand Prion Proteins for finding cure to various Human Brain Diseases. Building a XML Data Source based on PO will assist in discovery process.

- **ProteinOntology**
  - **AtomicBind**
  - **Atoms**
  - **Bind**
  - **Chains**
  - **Family**
  - **ProteinComplex**
    - **ChemicalBonds**
      - **CISPeptide**
      - **DisulphideBond**
      - **HydrogenBond**
      - **ResidueLink**
      - **SaltBridge**
    - **Constraints**
      - **GeneticDefects**
      - **Hydrophobicity**
      - **ModifiedResidue**
    - **Entry**
      - **Description**
      - **Molecule**
      - **Reference**
    - **FunctionalDomains**
      - **ActiveBindingSites**
      - **BiologicalFunction**
        - **PathologicalFunctions**
        - **PhysiologicalFunctions**
      - **SourceCell**
    - **StructuralDomains**
      - **Helices**
        - **Helix**
          - **HelixStructure**
      - **OtherFolds**
        - **Turn**
          - **TurnStructure**
      - **Sheets**
        - **Sheet**
          - **Strands**
      - **Structure**
        - **ATOMSequence**
        - **UnitCell**
  - **Residues**
  - **SiteGroup**

Figure 1: Class Hierarchy of Protein Ontology

2. Protein Ontology and OWL

As technologies mature, the shift from single annotation databases being queried by web-based scripts generating HTML pages to annotation repositories capable of exporting selected data in XML format, either to be further analysed by remote applications, or to undergo a transformation stage to be presented to user in a web browser – will undoubtedly be one of the major evolutions of protein annotation process. XML is a markup language much like HTML, but XML describes data using hierarchy. An XML document uses the schema to describe data and is designed to be self descriptive. This allows easy and powerful manipulation of data in XML documents. XML provides syntax for structured documents, but imposes no semantic constraints on the meaning of these documents.

Resource Description Framework (RDF) is a data model for objects or resources and relations between them, provides a simple semantics for this data model, and these data models can be represented in XML syntax. RDF Schema is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes.

To efficiently represent the protein annotation framework and to integrate all the existing data representations into a standardized protein data specification for the bioinformatics community, the protein ontology need to be represented in a format that not enforce semantic constraints on protein data, but can also facilitate reasoning tasks on protein data using semantic query algebra. This motivates the representation of Protein Ontology (PO) Model in Web Ontology Language (OWL). OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing additional vocabulary along with a formal semantics. Knowledge captured from protein data using OWL is classified in a rich hierarchy of concepts and their inter-relationships. OWL is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval and querying. We investigated the use of OWL for making Protein Ontology (PO) using Protégé OWL Plug-in.

OWL allows us to write explicit, formal concepts of describing protein data. Use of OWL to define formal protein data concepts provides: (1) well-defined syntax; (2) semantics, which is already present in protein data; (3) convenience of expression of integrated protein data using query algebra. Well-defined and structured syntax of protein ontology is necessary for machine processing and mining of protein data. Formal semantics describes the meaning of knowledge in protein data precisely. One of the uses of formal semantics is to allow people to reason about knowledge of protein domain. For the case of Protein Ontology, we may reason about:

- Class membership. If M is an instance of class Molecule, and Molecule is a subclass of Entry, then we can infer that M is an instance of Entry.
- Equivalence of classes. If Class HelixStructure is equivalent to class TurnStructure, and class TurnStructure is equivalent to class OtherFoldsStructure, then HelixStructure is equivalent to OtherFoldsStructure too.
- Classification. If we have declared that certain property-value pairs for Residue class should satisfy the condition that Residue should be a 3-

letter word for membership of Residue Class, then if an individual R satisfies such a condition, we can conclude R is instance of Residue Class.

3. PO Benefits and Limitations

Apart from classifying or organizing protein data and knowledge about proteins in a hierarchy, PO has following benefits:

1. Protein Ontology (PO) provides a unified vocabulary for capturing declarative knowledge about protein domain and to classify that knowledge. Information captured by PO is classified in a rich hierarchy of concepts and their inter-relationships.
2. In PO the notions classification, reasoning, and consistency are applied by defining new concepts or classes from defined generic concepts or classes. The concepts derived from generic concepts are placed precisely into class hierarchy of Protein Ontology to completely represent information defining a protein complex.

OWL fits to be used as development language for OWL as of following reasons:

1. As the OWL representation used in Protein Ontology is an XML-Abbrev based (Abbreviated XML Notation), it can be easily transformed to the corresponding RDF and XML formats without much effort using the available converters.
2. Most of the Other Biomedical Ontologies in Genetics and Molecular Biology are represented in OWL, such as: Gene Ontology (GO) [GO 2001], RiboWEB [Altman et al., 1999] and UMLS [UMLS 1993].

We are constantly working to improve PO features. Here are some of the improvements that we are looking at on achieving by next year:

1. For Protein Functional Classification, in addition to presence of domains, motifs or functional residues, following factors are relevant: (a) similarity of three dimensional protein structures, (b) proximity to genes (may indicate that proteins they produce are involved in same pathway), (c) metabolic functions of organisms and (d) evolutionary history of the protein. At the moment PO's Functional Domain Classification does not address the issues of proximity of genes and evolutionary history of proteins. These factors will be added in future to complete the Functional Domain Classification System in PO.
2. The Constraints defined in PO are not mapped back to protein sequence, structure and function they affect. Achieving this in future will inter-link all the concepts of PO.
3. We are in process of defining semantic query algebra for PO to efficiently reason and query the underlying XML database.
4. We will soon provide secured user interfaces to browse, query, and add protein data instances in PO.

4. Concluding Remarks

The overall objective of Protein Ontology (PO) Project is: “To correlate information about multiprotein machines with data in major protein databases to better understand sequence, structure and function of protein machines.” OWL provides a language for capturing declarative knowledge about protein domain and a classifier that allows reasoning about protein data. Knowledge captured from protein data using OWL is classified in a rich hierarchy of concepts and their inter-relationships. We investigated the use of OWL for making Protein Ontology (PO) using Protégé OWL Plug-in. OWL is flexible and powerful enough to capture and classify biological concepts of proteins in a consistent and principled fashion. OWL is used to construct Protein Ontology (PO) that can be used for making inferences from proteomics data using defined semantic query algebra.

References

[Altman et al. 1999] Altmann, R. B., M. Bada, et al. (1999). "RiboWeb: An Ontology-Based System for Collaborative Molecular Biology." IEEE Intelligent Systems (SEPTEMBER/OCTOBER 1999): 68-76.

[GO 2001] GO. (2001). "Creating the Gene Ontology Resource: Design and Implementation." Genome Research 11: 1425-1433.

[Sidhu et al., 2006] Sidhu, A. S., T. S. Dillon, et al. (2006). Protein Ontology Project: 2006 Updates (Invited Paper). Data Mining and Information Engineering 2006. A. Zanasi, C. A. Brebbia and N. F. F. Ebecken. Prague, Czech Republic, WIT Press.

[Sidhu et al., 2005a] Sidhu, A. S., T. S. Dillon, et al. (2005). Ontological Foundation for Protein Data Models. First IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2005), In conjunction with On The Move Federated Conferences (OTM 2005). Agia Napa, Cyprus, Springer-Verlag. Lecture Notes in Computer Science (LNCS).

[Sidhu et al., 2005b] Sidhu, A. S., T. S. Dillon, et al. (2005). Protein Ontology: Semantic Data Integration in Proteomics. 4th International Joint Conference of InCoB, AASBi and KSBI (BIOINFO2005). Busan, Korea.

[Sidhu et al., 2005c] Sidhu, A. S., T. S. Dillon, et al. (2005). Protein Ontology: Vocabulary for Protein Data. 3rd IEEE International Conference on Information Technology and Applications (IEEE ICITA 2005). Sydney, IEEE CS Press. Volume 1: 465-469.

[Sidhu et al., 2004a] Sidhu, A. S., T. S. Dillon, et al. (2004). A Unified Representation of Protein Structure Databases (Book Section). Biotechnological Approaches for Sustainable Development. M. S. Reddy and S. Khanna. India, Allied Publishers Pty. Ltd., India: 396 - 408.

[Sidhu et al., 2004b] Sidhu, A. S., T. S. Dillon, et al. (2004). Making of Protein Ontology (Invited Paper). 2nd Australian and Medical Research Congress 2004. M. Kavallaris. Sydney, National Health and Medical Research Council, Australian Government: 150-151.

[Sidhu et al., 2004c] Sidhu, A. S., T. S. Dillon, et al. (2004). Protein Knowledge Base: Making of Protein Ontology (Invited Paper). HUPO 3rd Annual World Congress 2004. R. A. Bradshaw. Beijing, China., American Society for Biochemistry and Molecular Biology. Vol. 3, No. 10 Oct. (Sup.): S262.

[UMLS 1993] McCray et al. (1993). Representing biomedical knowledge in the UMLS semantic network. In: Broering NC, editor. High-performance medical libraries: advances in information management for the virtual era. Westport (CT): Meckler; 1993. p. 31-44.