Structured Vocabularies for Proteins

A. S. Sidhu, MIEEE*, T. S. Dillon, FIEEE* and E. Chang MIEEE**

* Faculty of Information Technology, University of Technology Sydney, Australia
** School of Information Systems, Curtin University of Technology Perth, Australia

* {asidhu, tharam}@it.uts.edu.au
** Elizabeth.Chang@cbs.curtin.edu.au

Abstract: In this paper we introduce various vocabularies and definitions used for defining Protein Ontology. Protein Ontology provides the technical and scientific infrastructure and knowledge to allow description and analysis of relationships between various proteins. Protein Ontology uses relevant protein data sources of information like PDB, SCOP, and OMIM. Protein Ontology describes: Protein Sequence and Structure Information, Protein Folding Process, Cellular Functions of Proteins, Molecular Bindings internal and external to Proteins, and Constraints affecting the Final Protein Conformation.

Introduction

Our Protein Ontology [1, 2, 3, 4, and 5] provides a common structured vocabulary for researchers who need to share knowledge in proteomics domain. It consists of concepts (or type definitions), which are data descriptors for proteomics data and the relations among these concepts. Protein Ontology has (1) a hierarchical classification of concepts represented as classes, from general to specific; (2) a list of attributes related to each concept, for each class; and (3) a set of relations between classes to link concepts in ontology in more complicated ways then implied by the hierarchy, to promote reuse of concepts in the ontology. Concrete examples or Instances of each Concept are shown in the Protein Ontology. Each attribute of an Instance may have a corresponding value, whereas classes only specify that the attribute exists. Protein Ontology provides a structured vocabulary description for protein domains that can be used to describe cellular products in any organism. Protein Ontology Framework describes: (1) Protein Sequence and Structure Information, (2) Protein Folding Process, (3) Cellular Functions of Proteins, (4) Molecular Bindings internal and external to Proteins and (5) Constraints affecting the Final Protein Conformation. The Protein Ontology currently contains 92 concepts or classes, 261 attributes or properties and 17550 instances, including 17347 instances for Protein Atoms. The XML Representation of the Database of Human Prion Proteins based on the proposed Protein Ontology is available on the Protein Ontology Website. There are a total of 17550 instances for 10 of the 57 Major Prion Proteins in the Database for various Protein Concepts defined by the Protein Ontology.

Protein Ontology Conceptual Framework

The Main Class of Protein Ontology is ProteinOntology. For each Protein that is entered into the knowledge base of protein ontology, submission information is entered into ProteinOntology Class. ProteinOntologyID has format like "PO000000052". There are six subclasses of ProteinOntology, called Generic Classes that are used to define complex concepts in other Protein Ontology Classes: Residues, Chains, Atoms, AtomicBind, Bind, and SiteGroup. Concepts from these generic classes are reused in various other Protein Ontology Classes for definition of Class Specific Concepts. Details and Properties of Residues in a Protein Sequence are defined by instances of Residues Class. Instances of Chains of Residues are defined in Chains Class. All the Three Dimensional Structure Data of Protein Atoms is represented as instances of Atoms Class. Defining Chains, Residues and Atoms as individual classes has the benefit that any special properties or changes affecting a particular chain, residue and atom can be easily added. Data about binding atoms in Chemical Bonds like Hydrogen Bond, Residue Links, and Salt Bridges is entered into ontology as an instance of AtomicBind Class. Similarly the data about binding residues in Chemical Bonds like Disulphide Bonds and CIS Peptides is entered into ontology as an instance of Bind Class. All data related to site groups of the active binding sites of Proteins is defined as instances of SiteGroup Class. Representation of Instances of Residues and Chains of Residues are shown as follows:

```
<Residues>
        <Residue>LEU</Residue>
        <ResidueName>LEUCINE</ResidueName>
        <ResidueProperty>1-LETTER CODE: L; FORMULA:
        C6 H13 N1 O2; MOLECULAR WEIGHT:
        131.17</ResidueProperty>
</Residues>

<Chains>
        <Chain>D</Chain>
        <ChainName>CHAIN D</ChainName>
</Chains>
```

The Root Class for definition of Protein Complexes in the Protein Ontology is ProteinComplex. The Protein

Complex Definition defines one or more Proteins in the Complex Molecule. There are six main subclasses within ProteinComplex class: Entry, Structure, StructuralDomains, FunctionalDomains, ChemicalBonds, and Constraints. These classes define sequence, structure and chemical binds present in the Protein Complex.

Entry Class

Entry Class specifies the details of a Protein or a Protein Complex that is entered into the knowledge base of protein ontology. Protein Entry Details are entered into Entry as instances of SourceDatabaseID, SourceDatabaseName and SubmissionDate. These attributes describe the entry in the original protein data source from where it was taken. Entry has three subclasses: Description, Molecule and Reference.

Structure Class

Structure Class of Protein Ontology defines the concept of ATOMSequence, reusing the definitions of Chain and Residues. ATOMSequence instance is constructed using generic concepts of Chains, Residues, and Atoms. The reasoning is already there in the underlying relationships and hierarchy of Protein Data, as each Chain in a Protein represents a sequence of Residues, and each Residue is defined by a number of three dimensional atoms in the Protein Structure. Structure Class also stores the Unit Cell Details.

```
<ATOMSequence>
<ProteinOntologyID>PO0000000004</ProteinOntologyID>
<_ATOM_Chain>A</_ATOM_Chain>
<_ATOM_Residue>ARG</_ATOM_Residue>
<AtomID>364</AtomID>
<Atom>HE</Atom>
<ATOMResSeqNum>148</ATOMResSeqNum>
<X>-23.549</X>
<Y>3.766</Y>
<Z>-0.325</Z>
<Occupancy>1</Occupancy>
<TempratureFactor>0</TempratureFactor>
<Element>H</Element>
</ATOMSequence>
```

Structural Domains Class

Similarly, in structural domains class secondary structure elements of protein structure like helices, sheets, and short loops can also be represented using generic concepts of Chains and Residues. For instance the hierarchy used in a Helices Instance of Protein Ontology differentiates general information about the Helices and the Helix Structure comprising of Chains and Residue Sequences as follows:

```
<Helices>
<ProteinOntologyID>PO0000000002</ProteinOntologyID>
<_StrDomain_SuperFamily>HAMSTER</_StrDomain_SuperFa
mily>
<_StrDomain_Family>PRION
PROTEINS</_StrDomain_Family>
<HelixID>1</HelixID>
<HelixNumber>1</HelixNumber>
<HelixClass>Right Handed Alpha</HelixClass>
<HelixLength>10</HelixLength>
<HelixStructure>
<_Helix_Chain>A</_Helix_Chain>
<_Helix_InitialResidue>ASP</_Helix_InitialResidue>
<HelixInitialResidueSeqNum>144</HelixInitialResidueSeqNum>
<_Helix_EndResidue>ASN</_Helix_EndResidue>
<HelixEndResidueSeqNum>153</HelixEndResidueSeqNum>
</HelixStructure>
</Helices>
```

Other secondary structures like sheets and loops are represented using concepts of chains and residues in the similar way. The Sheet Structures in Proteins are composed of various Strands and is represented as follows using Protein Ontology.

Functional Domains Class

PO has the first Functional Domain Classification Model defined using FunctionalDomains Class using: (1) Data about Cellular and Organism Source in SourceCell subclass and (2) Data about Biological Functions of Protein in BiologicalFunction subclass and (3) Data about Active Binding Sites in Proteins in ActiveBindingSites subclass. Like StructuralDomain Class, SuperFamily and Family Instances of generic class Family are used for identifying the Protein Family in FunctionalDomain Class. SourceCell specifies biological or chemical source of each biological molecule (Defined by Molecule Class) in the Protein. Biological Functions of the Protein Complex are described in BiologicalFunction. BiologicalFunction has two children, PhysiologicalFunction and PathologicalFunction, and each of these has several children and grand children describing various corresponding functions. The third subclass of FunctionalDomains is ActiveBindingSites that has details about active binding sites in the Protein. Active Binding Sites are represented in our ontology as a collection of various Site Groups, defined in SiteGroup class. SiteGroup has details about each of the Residues and Chain that form the Binding Site. There can be a maximum of seven Site Groups in the ontology. A typical instance of Source Cell in FunctionalDomains is:

```
<Source>
<ProteinOntologyID>PO0000000009</ProteinOntologyID>
<SourceMoleculeID>1</SourceMoleculeID>
<OrganismScientific>HOMO SAPIENS</OrganismScientific>
<OrganismCommon>HUMAN</OrganismCommon>
<ExpressionSystem>ESCHERICHIA COLI;
BACTERIA</ExpressionSystem>
<ExpressionSystemVector>PLAMID</ExpressionSystemVector>
<Plasmid>PRSETB</Plasmid>
</Source>
```

Chemical Bonds Class

Again the various chemical bonds used to bind various substructures in a complex protein structure are defined using generic concepts of Bind and Atomic Bind. The Chemical Bonds that have Binding Residues reuse the generic concept of Bind. In defining the generic concept of Bind in Protein Ontology we again reuse the generic concepts of Chains and Residues. Similarly the Chemical Bonds that have Binding Atoms reuse the generic concept of AtomicBind. In defining the generic concept of AtomicBind we reuse the generic concepts of Chains, Residues and Atoms.

```
<CISPeptides>
<ProteinOntologyID>PO0000000003</ProteinOntologyID>
<_Bind_Chain_1>H</_Bind_Chain_1>
<_Bind_Residue_1>GLU</_Bind_Residue_1>
<BindResSeqNum_1>145</BindResSeqNum_1>
<_Bind_Chain_2>H</_Bind_Chain_2>
<_Bind_Residue_2>PRO</_Bind_Residue_2>
<BindResSeqNum_2>146</BindResSeqNum_2>
<AngleMeasure>-6.61</AngleMeasure>
<Model>0</Model>
</CISPeptides>
```

Constraints Class

Various constraints that affect final protein conformation are defined in Constraints class using ConstraintID and ConstraintDescription. The constraints described in Protein Ontology at the moment are: (1) Monogenetic and Polygenetic defects present in genes that are present in molecules making proteins in GeneDefects subclass, (2) Hydrophobicity properties in Hydrophobicity Class, and (3) Modification in Residue Sequences due to Chemical Environment and Mutations are entered in ModifiedResidue Class. Post-translational residue modifications comprises those amino acids that are chemically changed in such way that they could not be restored by physiological processes, and other rare amino acids that are translationally incorporated but for historical reasons are represented as modified residues. The RESID Database is the most comprehensive collection of annotations and structures for protein modifications. The current version of RESID maps post-translational modifications to both PIR and Swiss-Prot. Data in GeneDefects class is entered as instances of GeneDefects Class and is normally taken from OMIM Knowledgebase or scientific literature. A typical instance of a Constraint is:

```
<Constraints>
<ProteinOntologyID>PO0000000009</ProteinOntologyID>
<ConstraintID> 3</ConstraintID>
<ConstraintDescription> MODIFICATION OF RESIDUES DUE
TO GLYCOSYLATION</ConstraintDescription>
</Constraints>
```

Advantages and Limitations of PO

Advantages of PO

1. Protein Ontology (PO) provides a unified vocabulary for capturing declarative knowledge about protein domain and to classify that knowledge to allow reasoning. Information captured by PO is classified in a rich hierarchy of concepts and their inter-relationships. PO is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval and querying.
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.
3. 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.

Limitations of PO and Future Work

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.

Conclusions

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". The Proposed Protein Ontology is the first ever work to integrate protein data based on data semantics describing various phases of protein structure. Protein Ontology helps to understand structure, cellular function and the constraints that affect protein in a cellular environment. The attribute values in the Protein Ontology are not defined as text strings or as set of keywords. Most of the Values are entered as instances of Concepts defined in Generic Classes. We defined a Database of 10 Human Prion Proteins based on the defined Protein Ontology, gathering information about various Prion Proteins from major Protein Databases.

References

[1] Sidhu, A. S., T. S. Dillon, et al. (2006). **Ontology for Data Integration in Protein Informatics** in "Database Modeling in Biology: Practices and Challenges". Z. Ma and J. Y. Chen. Springer, Inc., USA **(In Press).**

[2] Sidhu, A. S., T. S. Dillon, et al. (2004). **A Unified Representation of Protein Structure Databases** in "Biotechnological Approaches for Sustainable Development". M. S. Reddy and S. Khanna. India Allied Publishers Pty. Ltd., India: 396 – 408**, ISBN 81-7764-669-9.**

[3] Sidhu, A. S., T. S. Dillon, et al. (2006). **Protein Ontology Project: 2006 Updates (Invited Paper).** Seventh International Conference on Data, Text and Web Mining and their Business Applications and Management Information Engineering (Data Mining and Information Engineering 2006). A. Zanasi, C. A. Brebbia and N. F. F. Ebecken. Prague, Czech Republic, WIT Press.

[4] 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).

[5] 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.