OWL, Proteins and Data Integration

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. In this paper we propose an approach to integrate protein information from various data sources by defining a 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. Details about Protein Ontology are available online at **http://www.proteinontology.info/**.

Keywords: Protein Ontology, Biomedical Ontologies, Knowledge Representation, Information Retrieval, Data Integration.

1 Protein Ontology Overview

We defined a Protein Ontology [1, 2, 3, 4, 5] that 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 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 is available online at **http://www.proteinontology.info/**. Complete Documentation about the class hierarchy of Protein Ontology is available at the website. The Class Diagram and UML Diagrams, depicting Protein Ontology are also available at the website. The Ontology is defined by Web Ontology Language (OWL) and the complete OWL file is also available online. 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

S. Zhang and R. Jarvis (Eds.): AI 2005, LNAI 3809, pp. 1158 – 1161, 2005.

Protein Ontology Website. There are a total of 17550 instances for all of the 10 Major Prion Proteins in the Database for various Protein Concepts defined by the Protein Ontology.

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. 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.

2 Protein Ontology Implementation

Notions of classification, reasoning, and consistency are applied in the making of Protein Ontology 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. 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. To understand the reuse of concepts in Protein Ontology, here are some of the examples. 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.

```
<ATOMSequence>
        <NumberResidues>244</NumberResidues>
        <chain>
        <ChainID>A</ChainID>
        <Description>A Chain</Description>
        <residue>
        <ResidueID>ILE</ResidueID>
        <Description>ILE Type</Description>
        <ATOM>
        <ATOMID>1</ATOMID>
        <ATOM>N</ATOM>
        <residueSeqnum>16</residueSeqnum>
        <x>60.749</x>
        <y>50.351</y>
        <z>75.583</z>
        <occupancy>1.00</occupancy>
        <temperatureFactor>15.71</temperatureFactor>
        <element>N</element>
        </ATOM>
  7 More Atoms…..
</residue> </chain> </ATOM Sequence>
```

Similarly Secondary Structure elements of Protein Structure like helices, sheets, and short loops can also be represented using generic concepts of Chains and Residues. 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.

```
<Helices>
        <Helix>
        <HelixNumber> 1 </HelixNumber>
        <HelixID> HA </HelixID>
        <HelixLength> 9 </HelixLength>
                <HelixStructure>
                <Chain>
                <ChainID> A </ChainID>
                <intialResidue>GLY</intialResidue>
                <intialResSeqNum>86</intialResSeqNum>
                <endResidue>GLY</endResidue>
                <endResSeqNum>96</endResSeqNum>
        </Chain></HelixStructure></Helix>
… and so on for other helices present in Protein
</Helices>
```

Other secondary structures like sheets and loops are represented using concepts of chains and residues in the similar way. 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. Various other Chemical Bonds in Proteins can be defined in similar way.

```
<ChemicalBonds>
        <ResidueLink>
                <AtomicBind1>
                <AtomicBindResSeqNum>391</AtomicResSeqNum>
                <AtomicBindATOM>MN</AtomicBindATOM>
                <AtomicBindResidue>MN</AtomicBindResidue>
                </AtomicBind1>
                <AtomicBind2>
                <AtomicBindResSeqNum>217</AtomicResSeqNum>
                <AtomicBindATOM>OE2</AtomicBindATOM>
                <AtomicBindResidue>GLU</AtomicBindResidue>
                <AtomicBindSymmetry>2565</AtomicBindSymmetry>
                </AtomicBind2>
        </ResidueLink>
… and so on for other chemical bonds in Protein
</ChemicalBonds>
```

3 Conclusion

The explosion of protein data led to increased efforts to logically represent, store and display knowledge. There have been several domains which have successfully created standardized templates for data, and their usefulness is apparent. Protein Ontology improves on these online protein data resources in number of ways. Firstly, it contains templates for all kinds of protein data that is need to understand proteins, their functionality and the proteomics process itself. Previously there is not such integrated and structured data representation format available. Secondly, majority of the values for many attributes unlike previously are not simply text strings, but has been entered into the ontology as instances of other concepts, defined by Generic Classes.

References

[1] Sidhu, A. S., T. S. Dillon, et al. (2005). Ontology-based Knowledge Representation of Protein Data. 3rd International IEEE Conference on Industrial Informatics, Perth, Australia, IEEE CS Press.

[2] Sidhu, A. S., T. S. Dillon, et al. (2005). Protein Ontology: Vocabulary for Protein Data. 3rd IEEE International Conference on Information Technology and Applications. Sydney, Australia, IEEE CS Press.

[3] Sidhu, A. S., T. S. Dillon, et al. (2004). Making of Protein Ontology. 2nd Australian and Medical Research Congress 2004 (Invited Speaker). M. Kavallaris. Sydney, National Heath and Medical Research Council: 151.

[4] Sidhu, A. S., T. S. Dillon, et al. (2004). Protein Knowledge Base: Making of Protein Ontology. HUPO 3rd Annual World Congress 2004 (Invited Speaker). R. A. Bradshaw. Beijing, China, American Society for Biochemistry and Molecular Biology. 3: S262.

[5] 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. Mumbai, India, Allied Publishers Pvt. Ltd.: 396-408.