Protein Ontology: Semantic Data Integration in Proteomics

Amandeep S. Sidhu[1], Tharam S. Dillon[2], Elizabeth Chang[3] and Baldev S. Sidhu[4]
[1, 2]*Faculty of Information Technology, University of Technology Sydney, Australia,*
e-mail: (asidhu, tharam)@it.uts.edu.au
[3]*School of Information Systems, Curtin University of Technology Perth, Australia,*
e-mail: Elizabeth.Chang@cbs.curtin.edu.au
[4]*Punjab State Education Department, India, e-mail: bsidhu@biomap.org*

ABSTRACT: The Protein Structural and Functional Conservation need a common language for data definition. With the help of common language provided by Protein Ontology the high level of sequence and functional conservation can be extended to all organisms with the likelihood that proteins that carry out core biological processes will again be probable orthologues. The structural and functional conservation in these proteins presents both opportunities and challenges. The main opportunity lies in the possibility of automated transfer of protein data annotations from experimentally traceable model organisms to a less traceable organism based on protein sequence similarity. Such information can be used to improve human health or agriculture. The challenge lies in using a common language to transfer protein data annotations among different species of organisms. First step in achieving this huge challenge is producing a structured, precisely defined common vocabulary using Protein Ontology. The Protein Ontology described in this paper covers the sequence, structure and biological roles of Protein Complexes in any organism.

KEYWORDS: Protein Ontology, Protein Informatics, Biomedical Ontologies, Biomedical Systems, Data Integration, Systems Biology.

1 INTRODUCTION

The accelerating availability of protein sequences and structures has transformed both the theory and practice of computational biology. Where once biologists characterized proteins by their diverse activities and abundance, all biologists now acknowledge that there is likely to be a single limited universe of proteins, many of which are conserved in most or all living cells. This recognition has fuelled a unification of protein databases; the information about shared proteins contributes to our understanding of all the diverse organisms that share them. Knowledge of biological understanding of protein in one organism can certainly illuminate, and often provide strong inference of its role in other organisms.

The Protein Structural and Functional Conservation need a common language for data definition. With the help of common language provided by Protein Ontology the high level of sequence and functional conservation can be extended to all organisms with the likelihood that proteins that carry out core biological processes will again be probable orthologues. The structural and functional conservation in these proteins presents both opportunities and challenges. The main opportunity lies in the possibility of automated transfer of protein data annotations from experimentally traceable model organisms to a less traceable organism based on protein sequence similarity. Such information can be used to improve human health or agriculture. The challenge lies in using a common language to transfer protein data annotations among different species of organisms. First step in achieving this huge challenge is producing a structured, precisely defined common vocabulary using Protein Ontology. The Protein Ontology described in this paper covers the sequence, structure and biological roles of Protein Complexes in any organism.

2 PROTEIN ONTOLOGY PROJECT

The Protein Ontology Project seeks to provide a set of structured vocabularies for protein domains that can be used to describe cellular products in any organism. The work includes modeling Protein Structure and Experimental Data. Protein Ontology Framework describes: (1) Protein Sequence and Structure Information, (2) Protein Folding Process, (3) Cellular Functions of Proteins and (4) Molecular Bindings internal and external to Proteins and (5) External Factors affecting Final Protein Conformation. In this paper we will also discuss the implementation strategy for the Protein Ontology Project. Protein Ontology Project provides a community resource using these vocabularies promoting the use of common protein data representation. The Goal of the Protein Ontology Project is to produce a dynamic, controlled data and query vocabulary that can be applied to all proteins even as the knowledge of protein roles in cells is accumulating and changing. The motivations behind proposing a Protein Ontology are:

1. Efforts in building consensus on data format using semantics inherent in various protein databases. This can be attained by creation of a data representation standard that defines physiological models at atomic and molecular level. The ability of the protein ontology to define such models for protein molecules and then the ability to model single cells will provide basic data necessary to model entire organs and organisms automatically.
2. Biologists in different specialties tend to use different languages for description of same data. They have their particular theories and models for their own data collection of the domain they are

working on. Protein Ontology is a unified data description model that covers all of the working domains.

3. The terms used to describe biomolecular data has different granularity depending on the level at which the abstractions or concepts in the domain and have different scope. Therefore, terms used in different contexts have different meaning. Defining Protein Ontology brings a consistent structured terminology for all biomolecular data.
4. For various Protein Databases there are different data models. It is the interfaces that provide interoperation and data exchange, but there are no interfaces to recognize integration and interactions between various data models and to exchange Data and Meta Data between them in consistent format. Protein Ontology does the Data Integration and Data Exchange between various existing Protein Data Models.

Our Protein Ontology [1, 2, 3, 4, 5, 6, 7, 8, 9] defines a common structured vocabulary for researchers who need to share knowledge in proteomics domain. It includes concepts (type definitions), which are data descriptors for proteomics data and the relations among these concepts. The Key features of Protein Ontology are (1) a hierarchical classification of concepts (classes) from general to specific; (2) a list of attributes for each class; and (3) a set of relations between classes to link concepts in ontology in more complicated ways then implied by underlying hierarchy. The Concepts have instances, which represent concrete examples of more abstract classes found in internal part of the hierarchy. Each attribute of an instance may have a corresponding value, whereas classes only specify that the attribute exists.

3 IMPLEMENTATION

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. Most of the UML Diagrams depicting Protein Ontology along with the class diagram are available at the Protein Ontology Website (http://proteinontology.info/)

There are six subclasses of ProteinOntology that are used to define complex concepts in other classes of ProteinOntology: Residues, Chains, Atoms, AtomicBind, Bind, and SiteGroup. Concepts from these subclasses are referenced 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 a Protein Complex 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 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.

```
<Entry>
<ProteinOntologyID>PO0000000007</ProteinOntologyID>
<_Entry_SuperFamily>HUMAN</_Entry_SuperFamily>
<_Entry_Family>PRION PROTEINS</_Entry_Family>
<SourceDatabaseID>1E1P</SourceDatabaseID>
<SourceDatabaseName>PROTEIN DATA
BANK</SourceDatabaseName>
<SubmissionDate>09-MAY-00</SubmissionDate>
<Title>HUMAN PRION PROTEIN VARIANT
S170N</Title>
<Authors>L.CALZOLAI, D.A.LYSEK, P.GUNTERT,
C.VON SCHROETTER, R.ZAHN, R.RIEK,
K.WUTHRICH</Authors>
<Experiment>NMR, 20 STRUCTURES</Experiment>
<Keywords>PRION PROTEIN</Keywords>
<CitationTitle>NMR STRUCTURES OF THREE
SINGLE-RESIDUE VARIANTS OF THE HUMAN PRION
PROTEIN</CitationTitle>
<CitationAuthors>L.CALZOLAI, D.A.LYSEK,
P.GUNTERT, C.VON SCHROETTER, R.ZAHN, R.RIEK,
K.WUTHRICH</CitationAuthors>
<CitationPublication>PROC.NAT.ACAD.SCI.USA</Citati
onPublication>
<CitationReference>V.  97 8340 2000</CitationReference>
<CitationReferenceNumbers>ASTM PNASA6 US ISSN
0027-8424</CitationReferenceNumbers>
</Entry>
```

Structure has Protein Sequence and Structure data for a Protein Entry. Structure has two subclasses: ATOMSequence and UnitCell. ATOMSequence consists of various chains of residue sequences present in the Protein. Each Chain is a sequence of singular residues. Each Residue or Chain may have distinct properties and functionality. Each Residue has a number of atoms linked to it, that define the three dimensional structure of Protein. Here in Structure, Residue is a sub property of Chain and ATOM is the sub property of Residue. The Containment relationship: Chain < Residue < ATOM still represents the hierarchy need for protein sequence and structure data, but also preserves individuality of the components.

```
<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 Folds and Domains defining Secondary Structures of Proteins are defined in StructuralDomains. SuperFamily and Family Instances of StructuralDomains are used for identifying the Protein Family. The subclasses of StructuralDomains are Helices, Sheets, and OtherFolds. Helix, which is a subclass of Helices, identifies the helix using HelixNumber, HelixID, HelixClass, and HelixLength Instances. Helix has a subclass HelixStructure gives the detailed composition of the helix. Sheets contain all the data about sheets present protein using its subclass Sheet. Sheet identifies individual sheets using SheetID and NumberStrands which represents the Number of Strands in the Sheet. Sheet has subclass called Strands that lists strands starting with one edge of the sheet and continuing to the spatial adjacent strand. Third Subclass of StructuralDomains, OtherFolds consists of loosely coupled folds. One of the most common folds of this category is short loop turns which connect other secondary structure segments, described in Turn subclass of OtherFolds. A Turn is identified by Instances of TurnNumber and TurnID. Turn has a subclass TurnStructure that defines the detailed composition of a Turn.

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

Protein Ontology 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.

Chemical Bonds in a Protein are defined using ChemicalBonds class. Various Chemical Bonds defined in ontology by respective subclasses are: DisulphideBond, CISPeptide, HydrogenBond, ResidueLink, and SaltBridge. As said earlier, the 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. The respective classes defining specific chemical bonds use Bind to define participating binding Residues and AtomicBind to define participating binding 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>
```

Last subclass of Protein Complex describes the constraints that affect final protein conformation. 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. Data in GeneDefects class is entered as instances of GeneDefects Class and is normally taken from OMIM database [10] or literature.

4 RESULTS & DISCUSSIONS

The Ontology is available on the internet: http://www.proteinontology.info/. The Class Diagram and UML Diagrams, and the documentation for Protein Ontology are available at the website. 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. 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. There are a total of 17550 instances for all of the 57 Major Prion Proteins in the Database for various Protein Concepts defined by the Protein Ontology. Protein Ontology describes the concepts of interest in protein complex mechanisms and proteomics process. The protein data source attributes are mapped to these defined concepts. Some of the information used while defining these Type Definitions is taken from PDB [11, 12, 13, 14], SCOP [15, 16], and OMIM [10] databases. Protein Ontology will be useful for standardizing protein data representation and browsing, but its real power comes from the fact that computer programs can be written to automatically extract and analyze data.

5 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). Creating a Protein Ontology Resource. 2005 IEEE Computational Systems Bioinformatics Conference. Stanford University, California, IEEE CS Press.

[2] 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.

[3] 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.

[4] Sidhu, A. S., T. S. Dillon, et al. (2005). The Protein Ontology Project: Structured Vocabularies for Proteins. Data Mining 2005. A. Zanasi, C.A.Brebbia and N. F. F. Ebecken. Skiathos, Greece, WIT Press, Southampton, UK.

[5] Sidhu, A. S., T. S. Dillon, et al. (2005). Protein Ontology Project. Fourth Indo-US Workshop on Mathematical Chemistry, Indo-US Workshop on Mathematical Chemistry Series 2005 (Invited Speaker). S. Basak. Bioinformatics Centre, University of Pune, India.

[6] 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.

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

[8] 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.

[9] Sidhu, A. S., T. S. Dillon, et al. (2004). An XML based semantic protein map. Data Mining 2004. A. Zanasi, N. F. F. Ebecken and C.A.Brebbia. Malaga, Spain, WIT Press, Southampton, UK. 10: 51-60.

[10] McKusick, V. A. (2000). Online Mendelian Inheritance in Man, OMIM. Baltimore, MD, Johns Hopkins University, National Center for Biotechnology Information, and National Library of Medicine.

[11] Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." Journal of Molecular Biology 112(3): 535-42.

[12] Weissiga, H. and P. E. Bourne (2002). "Protein structure resources." Biological Crystallography D58: 908-915.

[13] Westbrook, J., Z. Feng, et al. (2002). "The Protein Data Bank: unifying the archive." Nucleic Acid Research 30(1): 245-248.

[14] Bhat, T. N., P. E. Bourne, et al. (2001). "The PDB data uniformity project." Nucleic Acid Research 29(1): 214-218.

[15] Conte, L. L., B. Ailey, et al. (2000). "SCOP: a Structural Classification of Proteins database." Nucleic Acids Research 28(1): 257-259.

[16] Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures." Journal of Molecular Biology 247: 536-540.