The protein ontology project: structured vocabularies for proteins

A. S. Sidhu¹, T. S. Dillon¹, B. S. Sidhu² & E. Chang³
¹Faculty of Information Technology, University of Technology Sydney, Australia
²Punjab State Education Department, India
³School of Information Systems, Curtin University of Technology, Australia

Abstract

The rapid generation of accessible protein data sources has generated confusion over protein data representation. 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 modelling protein structure and experimental data. The 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) External factors affecting final protein conformation. In this paper we will also discuss the implementation strategy for the protein ontology project. The 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.

Keywords: protein ontology, proteomics, bioinformatics, protein informatics, computational proteomics, protein structure, biomedical ontologies, data integration, data semantics.

1 Introduction

Prediction of protein folding pathway may be evaluated by predicting subsegments or substructures of proteins. If computational model has right
underlying assumptions about what comes first in the pathway, and what comes next, and so on, then blind predictions such as those done as a part of protein structure assessment may validate that model. We define a protein ontology model that describes the concepts of interest in protein complex mechanisms and the protein data source characteristics are mapped to these concepts. The creation of a Protein Ontology that provides a comprehensive understanding of Protein Complex Mechanisms will help in the understanding of Cellular Mechanisms. Diverse types of data formats taken from different protein data sources are represented using a set of type definitions within this protein ontology, and these data are linked to each other with numerous connections. Not only does this structured representation allow easier data retrieval to users, but it also facilitates automated data mining by computer programs. In this paper, we describe the design principles behind the proposed Protein Ontology, illustrate how we have represented certain key data types to represent protein data, and describe the resulting Protein Ontology as it is currently publicly available.

Ontology & Knowledge Base approaches similar to the proposed approach like Gene Ontology (Harris et al. [7], Yeh et al. [16], and Ashburner et al. [3]) and RiboWEB (Altman et al. [2], Bada and Altman [14] and Abernethy et al. [1]) exist for Genes and RNA. The creation of a Protein Ontology (Sidhu [10–13]) that provides a comprehensive understanding of Protein Complex Mechanisms will completely map the understanding of Central Dogma. Protein Ontology will facilitate computational processing data, and develop standardized structured data representation model formats for proteomics data. It will make it possible to study relationships among proteins, protein folding, behavior of protein under various environments, and most importantly cellular function of protein.

2 Implementation

Protein Ontology 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 or properties 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. The attributes or properties of each class can be of object type or data type. Object Properties are used in Complex Concepts to refer to the definitions of Generic Classes. For example _Structure_Residue is an Object Property of ATOMSequence Class which fetches instance of Residue from Residues Class. Some of the information while defining these Type Definitions is taken from PDB (Weissiga [14], Westbrook et al. [15], and Bhat et al. [5]), SCOP (Conte et al. [6] and Murzin et al. [9]) and OMIM (McKusick [8]) databases.
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 “PO0000000007”.

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.

The Root Class for definition of a Protein Complex in the Protein Ontology is ProteinComplex. There are six main subclasses within ProteinComplex class: Entry, Structure, StructuralDomains, FunctionalDomains, ChemicalBonds, and Constraints.

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. Description has data about title of the entry, authors of the entry, experiment that produced the entry and keywords describing the entry. The second subclass of Entry is Molecule which is simply any chemically distinct molecule or compound in a protein complex. MoleculeID just uniquely identifies a Molecule. MoleculeName is the Chemical Name of the Molecule. Molecule Chain refers the Chain Description. BiologicalUnit Instance describes the larger biological unit of which molecule is a part. Engineered identifies whether the molecule is engineered using Recombinant Technology or Chemical Synthesis. A specific domain or region of the molecule is defined using Fragment. Mutated Molecules of the Protein have Mutations Information. Details about various mutations are described in GeneticDefects Class. List of Synonyms for Molecule Name are in Synonyms. OtherDetails describes any other information. Reference subclass lists the various literature citations of the protein or protein complex described by the instances of: CitationReference, CitationPublication, CitationTitle, CitationAuthors, CitationEditors, and CitationReferenceNumbers.
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. Data from Protein Crystallography like a, b, c, alpha, beta, gamma, z, and SpaceGroup are entered in UnitCell.

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 in terms of following instances:

- **Helix Chain**: Chain of Strand (References Chain Details from Chains Class).
- **Helix Initial Residue**: Initial Residue of each Helix (References Residue Details from Residues Class).
- **Helix Initial Residue Sequence Number**: Identifies the Residue Sequence Number of the Initial Residue in the Helix.
- **Helix End Residue**: End Residue of each Helix (References Residue Details from Residues Class).
- **Helix End Residue Sequence Number**: Identifies the Residue Sequence Number of the End Residue in the Helix.

Second Subclass of StructuralDomains, Sheets contains 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 in terms of following:

- **Strand Number**: Strand Number for each strand within the Sheet.
- **Strand Chain**: Chain of Strand (References Chain Details from Chains Class).
- **Strand Initial Residue**: Initial Residue of each Strand (References Residue Details from Residues Class).
- **Strand Initial Residue Sequence Number**: Identifies the Residue Sequence Number of the Initial Residue in the Strand.
- **Strand End Residue**: Initial Residue of each Strand (References Residue Details from Residues Class).
- **Strand End Residue Sequence Number**: Identifies the Residue Sequence Number of the End Residue in the Strand.
• **Strand Sense**: Sense of Strand with respect to the previous strand in the sheet.

• **Strand Current ATOM**: ATOM in Current Strand (References Atom Details from Atoms Class).

• **Strand Current Residue**: Residue in Current Strand (References Residue Details from Residues Class).

• **Strand Current Residue Sequence Number**: Identifies the Residue Sequence Number of the Current Residue in the Strand.

• **Strand Previous ATOM**: ATOM in Previous Strand (References Atom Details from Atoms Class).

• **Strand Previous Residue**: Residue in Previous Strand (References Residue Details from Residues Class).

• **Strand Previous Residue Sequence Number**: Identifies the Residue Sequence Number of the Previous Residue in the 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 in terms of following instances:

• **Turn Chain**: Chain of Turn (References Chain Details from **Chains** Class).

• **Turn Initial Residue**: Initial Residue of each Turn (References Residue Details from **Residues** Class).

• **Turn Initial Residue Sequence Number**: Identifies the Residue Sequence Number of the Initial Residue in the Turn.

• **Turn End Residue**: End Residue of each Turn (References Residue Details from **Residues** Class).

• **Turn End Residue Sequence Number**: Identifies the Residue Sequence Number of the End Residue in the Turn.

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. **SourceCell** specifies biological or chemical source of each biological molecule (Defined by **Molecule** Class) in the Protein. **SourceMoleculeID** uniquely identifies each biological molecule. **SourceSynthetic** indicates a chemically-synthesized source. **SourceMoleculeFragment** specifies a domain or fragment of the biological molecule. The property is equivalent to **MoleculeID** property in **Molecule** Class. **OrganismScientific** and **OrganismCommon** are the Scientific Name and Common Name of the Organism respectively. **Strain** describes the Strain of the Source and **Variant** identifies the variant. **CellLine** Identifies the line of cells used in the experiment. **Organ** defines an organized group of tissues for a specific function. **Tissue** in itself is an organized group of cells with common
function. Cell identifies a particular cell type and Organelle is an organized structure within a cell. Secretion identifies the secretion such as saliva or venom, from which molecule was isolated. CellularLocation identifies the location inside or outside the cell. Plasmid describes the plasmid containing the gene and Gene gives detailed description of the gene. ExpressionSystem is the system used to express recombinant macromolecules. SourceOtherDetails is used to enter any additional data about the source. 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. 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.

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 Atomic Bind to define participating binding Atoms.

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, (2) Hydrophobicity of Proteins, and (3) The Modification in Residues due to any Chemical Effect. Gene Defect Data is entered as instances of GeneDefects Class and is normally taken from OMIM database (McKusick [8]) or literature. Details about the Hydrophobicity properties are entered into Hydrophobicity Class. Any Modification in Residue Sequences due to Chemical Environment and Mutations are entered in ModifiedResidue Class.

3 Results

The Ontology is available at: http://www.proteinontology.info/. The Class Diagram and UML Diagrams for Protein Ontology are available at the website. The Ontology Currently contains 76 concepts or classes, 212 attributes or properties and 79 instances. The ontology is 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.
4 Discussion

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.

In future, we will provide more instances to validate the Protein Ontology. In long term, we would like to create data input software that can be used to transfer data from Protein Databases into Protein Ontology Knowledge Base.

References


