Creating a Protein Ontology Resource

Amandeep S. Sidhu[1],Member IEEE, Tharam S. Dillon[1],Fellow IEEE, Elizabeth Chang[2],Member IEEE
[1]*Faculty of Information Technology, University of Technology Sydney, Australia*
{asidhu, tharam}@it.uts.edu.au
[2]*School of Information Systems, Curtin University of Technology, Perth, Australia*
Elizabeth.Chang@cbs.curtin.edu.au

Abstract

Protein Data Integration approaches at the moment considers data sources as data repositories, but not as applications; which in turn may embody complex interactions with other data sources. Current approaches do not provide methods both for generic mapping protein data representation, depicting interactions in data it describes and for interfacing existing data. The proposed Protein Ontology shows the value of hierarchy and relationships present in proteomics data. The creation of a Protein Ontology provides understanding of diverse types of data like: (1) Protein Entry Details, (2) 3D Structural Representations of Proteins, (3) Structural Folds and domains conserved in proteins, (4) Functional Domains and Families created based on Physiological and Pathological Functions of Proteins, and (5) Various Constraints like Genetic Defects and Chemical Properties of Cell that affect Final Stable Molecular Structure of Protein. Protein Ontology describes the concepts of interest in protein complex mechanisms and proteomics process.

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

1. Introduction

The life sciences activities are commonly categorized as computational biology (such as proteomics and genomics) and as database development and exploitation of biological data banks of macromolecules – Proteins, RNA and DNA. Heterogeneity among various information sources is a major issue when extracting value from various distributed biological resources available. Biological Knowledge has to be comprised of multiple sources when answering queries. Information integration from multiple protein databases like PDB, SWISS-PROT, and PIR needs multi database query formations when answering user queries. Multiple databases may cover same data, but there focus might be different. The SWISS-PROT database provides Protein Sequence Information, PDB database provides Protein Structure Information, and PIR is mainly for cross referencing and linking various protein references. To answer data from these databases the data needs to be combined and represented in consistent fashion. While these data formats are useful for knowledge extraction on per – protein basis, they do not allow for efficient integration of all proteomics data relevant to a particular experiment, and they are certainly not provide all the knowledge needed for protein complexes. It is therefore quite difficult to create self-consistent models, and evaluate the compatibility of individual protein family data sets with these models.

We propose a Protein Ontology, showing the value of structured representations of proteomics data. 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.

2. Protein Ontology Overview

We defined a Protein Ontology [1, 2, 3, 4, 5, 6, 7, 8] 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 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. 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.

3. Implementation

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 91 *concepts* or classes, 248 *attributes* or properties and 99 instances.

4. 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. (2005). The Protein Ontology Project: Structured Vocabularies for Proteins. Data Mining 2005. Skiathos, Greece, WIT Press, UK.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.