# Protein Ontology: Vocabulary for Protein Data

Amandeep S. Sidhu, Member, IEEE,
Tharam S. Dillon, Fellow, IEEE
*Faculty of Information Technology, University of Technology Sydney, Australia*
(asidhu, tharam)@ it.uts.edu.au

Elizabeth Chang, Member, IEEE
*School of Information Systems, Curtin University of Technology Perth, Australia*
*Elizabeth.Chang@cbs.curtin.edu.au*

Baldev S. Sidhu
*Punjab State Education Department, India*
*bsidhu@biomap.org*

## Abstract

*These Huge amounts of Protein Structure Data make it difficult to create explanatory and predictive models that are consistent with huge volume of data. Difficulty increase when large variety of heterogeneous approaches gathers data from multiple perspectives. In order to facilitate computational processing data, it is especially critical to develop standardized structured data representation model formats for proteomics data. In this paper we describe a Protein Ontology Model for integrating protein databases and deduce a structured vocabulary for understanding process of protein synthesis completely. Proposed Protein Ontology Model provides biologists and scientists with a description of sequence, structure and functions of protein and also provides interpretation of various factors on final protein structure conformation. The Structured Vocabulary for Protein Data, describing Protein Ontology is composed of various Type Definitions for Protein Entry Details, Sequence and Structural Information of Proteins, Structural Domain Family of Protein, Cellular Function of Protein, Chemical Bonds present in the Protein, and External Constraints deciding final protein conformation. The Proposed Ontology Model will provide easier ways to predict and understand proteins.*

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

## 1. Introduction

Amino Acid Sequence provides important into structure of proteins, which in turn greatly facilitates the understanding of its biochemical and cellular function. Efforts to use computational methods in predicting protein structure based only on sequence information started 30 years ago [1, 2]. Only during last decade, the introduction of new computational techniques such as protein fold recognition and the growth of the sequence and structure databases due to modern high throughput technologies led to an increase in success rate of protein prediction methods, so that they can be used by molecular biologists. The computational assignment of three dimensional structures to newly determined protein sequences is becoming an important experiment in protein structure determination and in structural genomics [3]. The prediction methods aim to predict approximate three-dimensional models for proteins bearing no evident sequence similarity to any protein of known structure [4]. The assignment is carried out by searching a library of known structures, usually obtained from databases like PDB [5, 6, 7, 8], SWISS-PROT [9, 10], and PIR [11] for a compatible fold. A variety of fold-recognition methods have been published, both structure dependent [12, 13, 14] and sequence only dependent [15, 16].

All computational models that predict something have certain underlying assumptions that constitute the physical basis for the model. In protein structure prediction, there are two physical/biological processes that can be modeled: the process of evolution and process of folding. The two paradigms governing these processes are Darwin and Boltzmann, named after scientists who defined principles of evolutionary biology and thermodynamics. Most of the work in protein structure prediction is Darwin-based, using premise that sequences that have common ancestor have similar folds. Most of the methods that use multiple sequence alignment, structural alignment, or "threading potentials" are implicitly searching for a common ancestor. Despite the "energy-like" scoring functions, these methods do not address the physical process of folding. Evolution happens in millions of years, whereas folding happens in fractions of a second. Protein structure prediction of Boltzmann kind is perceived to be very difficult problem. Many have failed over last thirty years, either to use Boltzmann-based prediction method or improve on exiting Darwin-based prediction methods. Data Explosion of Protein Structure Data and lack of availability of a vocabulary covering both data and semantic representations makes it difficult to create explanatory and predictive models that solve protein folding problem. Difficulty increase when large variety of heterogeneous approaches gathers data from multiple perspectives. In this paper we describe a structured vocabulary for understanding process of protein synthesis completely. The proposed vocabulary provides biologists and scientists with a description of sequence, structure and functions of protein and also provides interpretation of various factors on final protein structure conformation.

## 2. Need for Structured Vocabulary

Prediction of protein folding pathway may be evaluated by predicting sub-segments 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. For correctly defining assumptions and completely understanding processes of Protein Synthesis usually both data and its biological context determines the complete meaning (or semantics) of the protein structure. 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 arising need for data source transparency lead researchers to consider semantic integration [17, 18]. Karp [19, 20] has identified the several approaches that have been proposed and implemented by bioinformatics researchers and proposed a strategy for data interoperation. The Overall Objective (Goal 1, Aim 3 of DoE GTL [21]) of the Research is "To correlate information about multiprotein machines with structural information generated in NIH Protein Structure Initiative and other major Protein Databases to better understand the geometry, organization and function of protein machines". The objective can be achieved to some extent by creating a Protein Ontology [22, 23, 24, 25, 26, 27] for integrating protein databases and deducing a structured vocabulary for understanding process of protein synthesis completely. The Design Goals of proposed Protein Ontology are:

1. To compile a comprehensive structured vocabulary of terms describing various elements of proteins those are shared among life forms.
2. The terms are defined closely to Protein Data Bank, largest protein source available are organized into broader and narrow refinements.
3. The Vocabulary is cross referenced to various database schemas it integrates and to Unified Medical Language System (UMLS) Thesaurus to have cross validation of the context usage and have exact linkages among terms.
4. To describe various proteins in various organism models using these terms.

Ontology & Knowledge Base approaches similar to the proposed approach like Gene Ontology [28, 29, 30] and RiboWEB [31, 32] exist for Genes and RNA. The creation of a Protein Ontology 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, behaviour of protein under various environments, and most importantly cellular function of protein.

## 3. Type Definitions

The Structured Vocabulary for Protein Data (as in Figure 1) is composed of various Type Definitions for Protein Entry Details, Sequence and Structural Information of Proteins, Structural Domains of Protein, Functional Domains of Protein, Chemical Bonds present in the Protein, and Constraints deciding final protein conformation. Now let's briefly describe various types that make the proposed protein ontology.

### *A. Entry Type Definition*

Type Definition for Entry describes: (1) General Protein Entry Description in Description Class, (2) Information about Molecules present in Protein in Molecule Class and (3) Information about Citations of Protein Structures in Literature in References Class.

*B. Structure Type Definition*

Type Definition for Structure describes: (1) Protein Sequence & Structure information using concept of "ATOM Sequence" in ATOMSequence Class, and (2) Unit Cell Information in Unit Cell Class. We defined the concept of ATOMSequence from the following observation of representation of sequence and structural data of the proteins: *ATOMSequence consists of various chains of residue sequences present in the Protein. Each Chain is a sequence of singular residues, each having distinct properties and functionality. Each Residue has a number of atoms linked to it, that define the three dimensional structure of Protein.* Defining Chain, Residue and ATOM as individual classes has the benefit that any special properties or changes affecting a particular chain, residue and ATOM can be easily added. The Containment relationship: ATOM Sequence < Chain < Residue < ATOM still represents the hierarchy need for protein data representation, but also preserves individuality of the components.

*C. Structural Domains Type Definition*

Type Definition for Structural Domains describes the structural domains present in the Secondary Structure of Protein as: (1) All Helices defined by Helices Class, (2) All Sheets defined by Sheets Class, (3) All the loosely coupled folds defined by Other Folds Class. The Helices referenced in Helices Class are defined in the Helix Class in Detail. Similarly, the Sheets referenced in Sheets Class are defined in the Sheet Class in Detail. The Other Folds defined at the moment in Protein Ontology is short loops and turns defined in Turn Class.

*D. Functional Domains Type Definition*

Protein Ontology has the first Functional Domain Classification Model defined using FunctionalDomains Class using: (1) Data about Cellular and Organism Source in SourceCell Class, (2) Data about Biological Functions of Protein in BiologicalFunction Class and (3) Data about Active Binding Sites in Proteins in ActiveBindingSites Class.

*E. Chemical Bonds Type Definition*

Chemical Bonds in a Protein are defined using ChemicalBonds. Various Chemical Bonds defined in ontology by following classes: DisulphideBond, CISPeptide, HydrogenBond, ResidueLink, and SaltBridge. The binding atoms in Chemical Bonds like Hydrogen Bond, Residue Links, and Salt Bridges are entered into ontology as an instance of AtomicBind Class. Similarly the 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.

*F. Constraints Type Definition*

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 [35] or literature.

## 4. Results

The Ontology is available on the internet: **http://www.proteinontology.info/**. The Class Diagram and UML Diagrams for Protein Ontology are available at the website. The Ontology Currently contains 91 *concepts* or classes, 246 *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.

## 5. Discussion

Some of the information while defining these Type Definitions is taken from PDB [5, 6, 7, 8], SCOP [33, 34], and OMIM [35] databases. 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.

## 6. References

[1] Nagano, K. (1973). "Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure." Journal of Molecular Biology 75(2): 401-420.
[2] Chou, P. Y. and G. D. Fasman (1974). "Prediction of protein conformation." Biochemistry 13(2): 222 - 245.
[3] Fischer, D., D. Baker, et al. (2001). "We need both computer models and experiments." Nature 409(6820): 558.
[4] Cymerman, I. A., M.Feder, et al. (2004). "Computational Methods for Protein Structure Prediction and Fold Recognition." Nucleic Acids and Molecular Biology 15: 1-21.
[5] Weissiga, H. and P. E. Bourne (2002). "Protein structure resources." Biological Crystallography D58: 908-915.
[6] Westbrook, J., Z. Feng, et al. (2002). "The Protein Data Bank: unifying the archive." Nucleic Acid Research 30(1): 245-248.
[7] Bhat, T. N., P. E. Bourne, et al. (2001). "The PDB data uniformity project." Nucleic Acid Research 29(1): 214-218.
[8] 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.
[9] Bairoch, A. and R. Apweiler (1997). "The SWISS-PROT protein sequence data bank and its supplement TrEMBL." Nucleic Acids Research 25(1): 31–36.
[10] Bairoch, A., P. Bucher, et al. (1997). "The PROSITE database, its status in 1997." Nucleic Acid Research 25(1): 217–221.
[11] George, D. G., R. J. Dodson, et al. (1997). "The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database." Nucleic Acids Research 25(1): 24–27.
[12] Jones, D. T., W. R. Taylort, et al. (1992). "A new approach to protein fold recognition." Nature 358(6381): 86 - 89.
[13] Rost, B. (1995). TOPITS: threading one-dimensional predictions into three-dimensional structures. Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, United Kingdom, AAAI Press.
[14] Fischer, D. (2000). Hybrid Fold Recognition: Combining Sequence Derived Properties with Evolutionary Information. Pacific Symposium on Biocomputing, Hawaii, World Scientific.
[15] Karplus, K., C. Barrett, et al. (1999). "Predicting protein structure using only sequence information." Proteins: Structure, Function, and Genetics 37(Supplement 3): 121 - 125.
[16] Rychlewski, L., L. Jaroszewski, et al. (2000). "Comparison of sequence profiles. Strategies for structural predictions using sequence information." Protein Science 9(2): 232-241.
[17] Adak, S., V. Batra, et al. (2002). "Bioinformatics for Microarrays."
[18] Goble, C., R. Stevens, et al. (2001). "A Transparent Access to Multiple Bioinformatics Information Sources." IBM Systems Journal 40(2): 532-551.
[19] Karp, R. M. (2003). Keynote Address: The Role of Algorithmic Research in Computational Genomics. IEEE Computational Systems Bioinformatics (CSB'03).
[20] Karp, P. D. (2000). "An Ontology for Biological Function Based on Molecular Interactions." Bioinformatics 16(2).
[21] DoE (2001). Genomes to Life Accelerating Biological Discovery, U.S. Department of Energy: April 2001.
[22] Sidhu, A. S., T. S. Dillon, et al. (2005). "Protein Ontology Project". Fourth Indo-US Workshop on Mathematical Chemistry with Applications to Drug Discovery, Environmental Toxicology, Cheminformatics and Bioinformatics, Indo-US Workshop on Mathematical Chemistry Series 2005 (Invited Speaker), Bioinformatics Centre, University of Pune, India, International Society of Mathematical Chemistry.
[23] Sidhu, A. S., T. S. Dillon, et al. (2005). "The Protein Ontology Project: Structured Vocabularies for Proteins." Data Mining 2005, Greece, Wessex Institute of Technology (WIT), UK.
[24] Sidhu, A. S., T. S. Dillon, et al. (2004). Making of Protein Ontology. 2nd Australian and Medical Research Congress 2004 (Invited Speaker), Sydney, National Heath and Medical Research Council.
[25] Sidhu, A. S., T. S. Dillon, et al. (2004). Protein Knowledge Base: Making of Protein Ontology. HUPO 3rd Annual World Congress 2004, Beijing, China, American Society for Biochemistry and Molecular Biology.
[26] Sidhu, A. S., T. S. Dillon, et al. (2004). A Unified Representation of Protein Structure Databases. Bioconvergence 2004 (Invited Paper), Punjab, India, 145.
[27] Sidhu, A. S., T. S. Dillon, et al. (2004). An XML based semantic protein map. Data Mining 2004, Malaga, Spain, WIT Press.
[28] Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Research 32(Database issue): 258-261.
[29] Yeh, I., P. D. Karp, et al. (2003). "Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO)." Bioinformatics 19(2): 241-248.
[30] Ashburner, M., C. A. Ball, et al. (2001). "Creating the Gene Ontology Resource: Design and Implementation." Genome Research 11: 1425–1433.
[31] Altman, R. B., M. Bada, et al. (1999). "RiboWeb: An Ontology-Based System for Collaborative Molecular Biology." IEEE Intelligent Systems (September/October 1999): 68-76.
[32] Bada, M. A. and R. B. Altman (1999). Computational Modeling of Structured Experimental Data. Stanford, CA, Stanford Medical Informatics SMI-1999-0764.
[33] Conte, L. L., B. Ailey, et al. (2000). "SCOP: a Structural Classification of Proteins database." Nucleic Acids Research 28(1): 257-259.
[34] 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.
[35] 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.

## 7. Figures

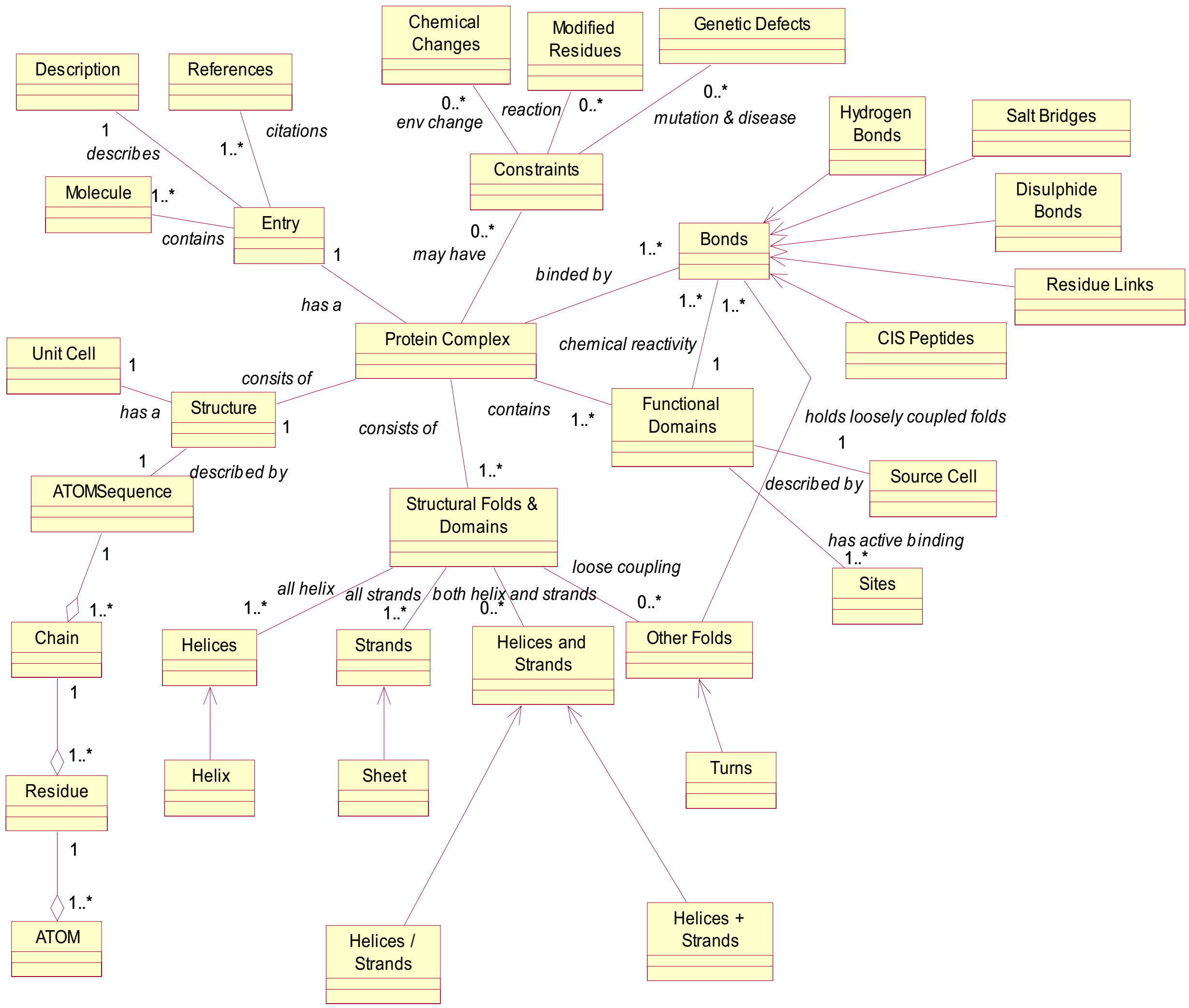

**Figure 1: Complete Protein Ontology Data Model**