Ontological Foundation for Protein Data Models

Amandeep S. Sidhu[1], Tharam S. Dillon[1], and Elizabeth Chang[2]

[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

Abstract. In this paper we proposed a Protein Ontology to integrate protein data and information from various Protein Data Sources. 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. We also created a database of 10 Major Prion Proteins available in various Protein data sources, based on the vocabulary provided by Protein Ontology. Details about Protein Ontology are available online at http://www.proteinontology.info/.

1 Introduction

A large variety of proteins have been deduced from the various genome projects and within them have been identified conserved or variant regions, functional and structural elements, features and domains. The continuum of life forms has become clearer and differences between species measurable. Novel biocatalysts and parameters relating structure to function have been identified from diversity of living organisms, and the network of molecular interactions and complex biological processes has become available for modelling. The primary tools for drug discovery are now the classification of chemical compounds, proteins and targets into functional groups, identification of relations between distant chemical targets & cells, drug effects on cells and lastly the knowledge of chemical structures and properties. Advanced protein engineering with the help of computer science has proven a sophisticated aid in the development of new biocatalysts, therapeutics and diagnostic tools. Advanced methods like high-throughput crystallography and nuclear magnetic resonance (NMR) have accelerated the resolution of new protein structures, and the modelling of macromolecules have been improved new groups of 3D protein structures.

These developments increased the importance of proteomics in heath care. Protein Informatics is a multidisciplinary field that is a synergy of proteomics process, bioinformatics and computational biology. The main objective of Protein Informatics

R. Meersman et al. (Eds.): OTM Workshops 2005, LNCS 3762, pp. 916 – 925, 2005.

is to provide a framework for developing, integrating and sharing knowledge present in various protein data sources. The advances in information and communication technologies coupled with increased knowledge about genes and proteins have opened new perspectives for study of protein complexes. There is a growing need to integrate the knowledge about various protein complexes for effective disease prevention mechanisms, individualized medicines and treatments and other accepts of healthcare. The exploitation of data from bioinformatics, medical informatics, medical imaging and clinical data requires a new and synergetic approach that enables a bilateral dialogue between these scientific disciplines, and integration in terms of data, methods, technology, tools and applications.

The proposed Protein Ontology provides the technical and scientific infrastructure and knowledge to allow evidence based description and analysis of relationships between proteins and other macromolecules. Protein Ontology uses all relevant protein data sources of information. The sources include new proteome information resources like PDB [1, 2, 3, 4] and SCOP [5, 6], as well as classical sources of information where information is maintained in a knowledge base of scientific text files like OMIM [7] and from various published scientific literature in various journals. PDB [1, 2, 3, 4] mainly provides protein entry and structure information for the Protein Ontology. SCOP [5, 6] provides the structural domain classification system that is widely used in proteomics. OMIM [7] is a knowledge base that provides texts and literature on various gene defects affecting the genes that make proteins. The information about the newly proposed functional domain classification system is gathered from various scientific literature and texts. On the whole Protein Ontology is an effort to seamlessly integrate all the data and knowledge about Proteins, to provide a data specification for new data representation and existing mining of existing data.

2 Ontology Foundations

Traditionally the knowledge base in biology has resided within the heads of experienced biologists and scientists who devoted study and time to become experts in their particular domain. This approach worked well in past when considerable effort was needed to tease data out of biological experiments, the flow of data was not so great to overwhelm the expert. However this situation is rapidly changing, many protein complexes are appearing each year and new experimental techniques are providing information on protein interactions. Not only is the rate of data acquisition growing exponentially but also a single experiment can collect data on huge range of molecules that would need many domain experts to interpret. There is therefore a need to create systems that can apply knowledge of domain experts to biological data. It is not envisaged that such systems will perform better than human experts; however they could play a crucial role in filtering the flood of data to point where human experts could again apply their knowledge. This then raises the questions, in particular how concepts and their relationships can be captured in ways that make them computationally available and traceable.

Ontology is a system that describes concepts and the relationships between them. Therefore, we proposed to build ontology for the Proteomics Domain. It is important to point out that this is an integration of data formats for representation. The Protein Ontology is an ontology based integration of heterogenous protein and biological data sources. Protein Ontology converts the enormous amounts of data collected by geneticists and molecular biologists into information that scientists, physicians and other health care professionals and researchers can use to easily understand the mapping of relationships inside protein molecules, interaction between two protein molecules and interactions between protein and other macromolecules at cellular level. Protein Ontology also helps to codify proteomics data for analysis by researchers.

A considerable body of research in the area of knowledge representation has shown that ontology must necessarily reflect a specific view of the data. Traditionally, ontologies have been represented using static models [8]. These can assist in exchanging knowledge at a purely terminological or syntactic level, but can suffer due to difficulties of interpretation; the relationships in the model rely solely on the prospective of the modeller. If we are to share knowledge, a clearer semantics is required. Full interaction with ontology requires, in addition a notion of functionality or reasoning the ontology can provide. Frame representations provide a precise, definitional framework to capture concepts and relationships between them. Frame formalism has been used to model biological data in EcoCyc Encyclopedia of E.Coli genes and metabolism [9]. The representation is however, static in the sense that kind-of hierarchy is asserted by the modeller, rather than deduced by the system from the descriptions of the concepts.

Description Logics (DL) [10] is an example of knowledge representation language. DL provides a language for capturing declarative knowledge about a domain and a classifier that allows reasoning about that knowledge. Information captured using DL is classified in a rich hierarchy of concepts and their inter-relationships. DL is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval and querying. In this paper we investigated use of OWL Description Logics (OWL-DL) for making Protein Ontology using Protégé OWL Plug-in. The OWL-DL is chosen for its following features:

1. OWL-DL is flexible and powerful enough to capture and classify biological concepts of proteins in a consistent and principled fashion.
2. OWL-DL is used to construct protein ontology that can be used for making inferences from proteomics data.

3 Related Work

In this section we will discuss various biomedical ontology works related to Protein Ontology. Gene Ontology (GO) [11] defines a hierarchy of terms related to genome annotation. GO is a structured network consisting of defined terms and relationships that describe Molecular Functions, Biological Processes, and Cellular Components of Genes. GO is clearly defined and modelled for numerous other biological ontology

projects [12]. So far GO has been used to describe the genes of several model organisms *(Saccharomyces cerevisiae, Drosophila melanogaster, Mus musculus and others)*.

RiboWEB [13] is an online data resource for Ribosome, a vital cellular apparatus. It contains a large knowledge base of relevant published data and computational modules that can process this data to test hypotheses about ribosome's structure. The system is built around the concept of ontology. Diverse types of data taken principally from published journal articles is represented using a set of templates in the knowledge base, and the data is linked to each other with numerous connections.

Protein Data Bank (PDB) has recently released versions of the PDB Exchange Dictionary and the PDB archival files in XML format collectively named PDBML [14]. The representation of PDB data in XML builds from content of PDB Exchange Dictionary, both for assignment of data item names and defining data organization. PDB Exchange and XML Representations use same logical data organization. A side effect of maintaining a logical correspondence with PDB Exchange representation is that PDBML lack the hierarchical structure characteristic of XML data.

PRONTO [15] is a directed acyclic graph (DAG) based ontology induction tool that constructs a protein ontology including protein names found in MEDLINE abstracts and in UNIPROT. It is a typical example typical example of mining literature and data sources. It can't be classified as protein ontology as it only represents relationship between protein literatures and does not formalize knowledge about protein synthesis process. Ontology for Protein Domain must contain terms or concepts relevant to protein synthesis, describing Protein Sequence, Structure and Function and relationships between them. While defining PO we made an effort to emulate the protein synthesis and describe concepts and relationships describing it.
There is a need for more agreed-upon semantical standard to describe protein data. PO addresses this issue by providing clear and unambiguous definitions of all major biological concepts of protein synthesis process and relationship between them using OWL. The use OWL in PO provides a unified controlled vocabulary both for annotation data types and for annotation data.

4 Protein Ontology Overview

We defined a Protein Ontology (PO) [16, 17, 18, 19, 20] that provides a common structured vocabulary for researchers who need to share knowledge in proteomics domain. PO consists of concepts (or classes), which are data descriptors for proteomics data and the relations among these concepts. PO 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. At the moment PO currently contains 92 *concepts* or classes and 261 *attributes* or properties. The structure of PO provides the concepts necessary to describe individual proteins, but does not contain individual protein themselves. The underlying XML database based on PO acts as

instance store for the PO. The Database consists of 17550 instances of 10 major prion proteins for various concepts defined in PO. PO 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 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 "PO0000000005". There are seven subclasses of ProteinOntology (PO), called Generic Classes that are used to define complex concepts in other PO Classes: Residues, Chains, Atoms, Family, AtomicBind, Bind, and SiteGroup. Concepts from these generic classes are reused in various other PO 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. Protein Family class represents Protein Superfamily and Family Details of Proteins. 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 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, function, and chemical bindings present in the Protein Complex. The Complete Class Hierarchy of Protein Ontology (PO) is shown in Figure 1**.** More detailed UML Diagrams for PO are available at the website.

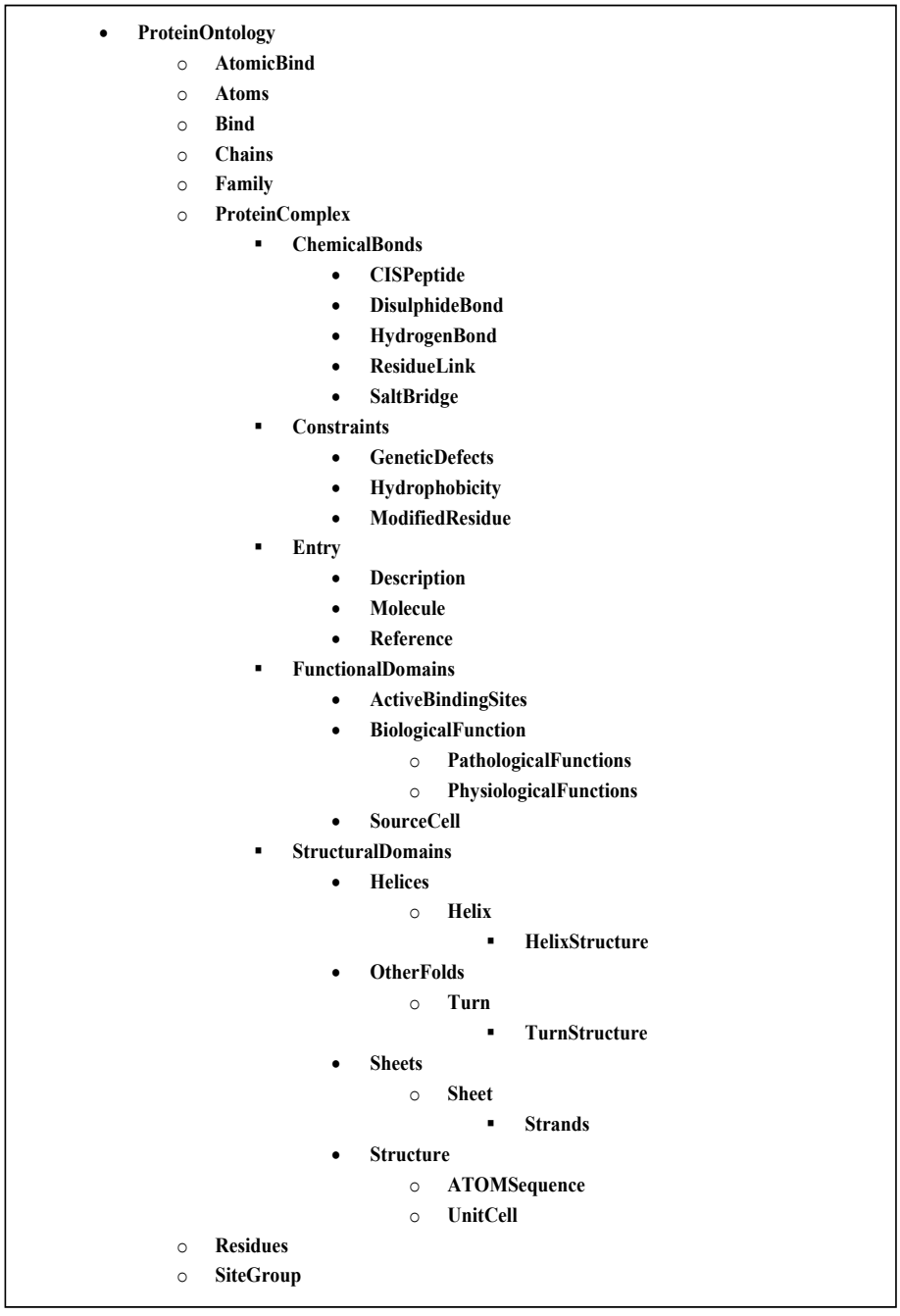

Fig. 1. Class Hierarchy of Protein Ontology

5 Protein Ontology Implementation

OWL-DL relies on notions classification, reasoning, and consistency. These notions 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. The Structure Class of Protein Ontology that is used to define ATOMSequence, with references to definitions of Chain and Residues.

```
<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>
```

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>
        <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>
        <HelixStructure>
```

```
            <_Helix_Chain>A</_Helix_Chain>
            <_Helix_InitialResidue>ASP</_Helix_InitialResidue>
            <HelixInitialResidueSeqNum>144</HelixInitialResidueSeqNum>
            <_Helix_EndResidue>ASN</_Helix_EndResidue>
            <HelixEndResidueSeqNum>153</HelixEndResidueSeqNum>
        </HelixStructure>
</Helices>
```

Other secondary structures like sheets and loops are represented using concepts of chains and residues in the similar way. The Sheet Structures in Proteins are composed of various Strands and is represented as follows using Protein Ontology.

```
<Sheets>
            <ProteinOntologyID>PO0000000001</ProteinOntologyID>
            <_StrDomain_SuperFamily>MOUSE</_StrDomain_SuperFamily>
            <_StrDomain_Family>PRION PROTEINS</_StrDomain_Family>
            <SheetID>S1</SheetID>
            <NumberStrands>2</NumberStrands>
            <Strands>
                <StrandNumber>2</StrandNumber>
                <_Strand_Chain>NULL</_Strand_Chain>
                <_Strand_IntialResidue>VAL</_Strand_IntialResidue>
                <StrandIntialResidueSeqNum>161</StrandIntialResidueSeqNum>
                <_Strand_EndResidue>ARG</_Strand_EndResidue>
                <StrandEndResidueSeqNum>164</StrandEndResidueSeqNum>
                <StrandSense>ANTI-PARALLEL</StrandSense>
</Strands></Sheets>
```

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.

```
<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>
```

A XML Database of 10 Major Prion Proteins available in various Protein data sources, based on the vocabulary provided by Protein Ontology is available on the PO

website. Soon we will have all the 57 Prion Proteins known to exist, and user interfaces to browse and query the database. The XML database currently contains 24 tables, 261 attributes and 17550 instances. 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. Building a XML Data Source based on PO will assist in discovery process.

6 Conclusion

Protein Ontology (PO) provides a unified vocabulary for capturing declarative knowledge about protein domain and to classify that knowledge to allow reasoning. Information captured by PO is classified in a rich hierarchy of concepts and their inter-relationships. PO is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval and querying. In PO the notions classification, reasoning, and consistency are applied 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 complex. 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. Our Protein Ontology (PO) is the first ever work to integrate protein data based on data semantics describing various phases of protein structure. PO helps to understand structure, cellular function and the constraints that affect protein in a cellular environment. The attribute values in the PO are not defined as text strings or as set of keywords. Most of the Values are entered as instances of Concepts defined in Generic Classes. We defined a XML Database of Human Prion Proteins based on PO, gathering information about various Prion Proteins from various data sources. For Protein Functional Classification, in addition to presence of domains, motifs or functional residues, following factors are relevant: (a) similarity of three dimensional protein structures, (b) proximity to genes (may indicate that proteins they produce are involved in same pathway), (c) metabolic functions of organisms and (d) evolutionary history of the protein. At the moment PO's Functional Domain Classification does not address the issues of proximity of genes and evolutionary history of proteins. These factors will be added in future to complete the Functional Domain Classification System in PO. The Constraints defined in PO are not mapped back to protein sequence, structure and function they affect. Achieving this in future will inter-link all the concepts of PO.

References

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

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

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

[4] 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.

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

[6] 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.

[7] 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.

[8] Schulze-Kremer, S. (1998). "Ontologies for Molecular Biology." Proceedings of Third Pacific Symposium on Biocomputing, Hawaii, AAAI Press: 693-704.

[9] Karp, P., M. Riley, et al. (1998). "Ecocyc: Electronic Encyclopedia of E.coli genes and metabolism." Nucleic Acids Research 26, 50.

[10] Borgida, A. (1995) "Description Logics in Data Management." IEEE Transactions on Knowledge and Data Engineering, 7, 671-682.

[11] GO. (2001). "Creating the Gene Ontology Resource: Design and Implementation." Genome Research 11: 1425-1433.

[12] GO. (2004). "Gene Ontology: looking backwards and forwards." Genome Biology 6(1): 103.1-103.4.

[13] Altmann, R. B., M. Bada, et al. (1999). "riboweb: An Ontology-Based System for Collaborative Molecular Biology." IEEE Intelligent Systems (SEPTEMBER/OCTOBER 1999): 68-76.

[14] Westbrook, J., N. Ito, et al. (2005). "PDBML: The Representation of Archival Macromolecular Structure Data in XML." Bioinformatics 21(7): 988-992.

[15] Mani, I., Z. Hu, et al. (2004). PRONTO: A Large-scale Machine-induced Protein Ontology. 2nd Standards and Ontologies for Functional Genomics Conference (SOFG 2004), UK.

[16] Sidhu, A. S., T. S. Dillon, et al. (2005). An Ontology for Protein Data Models. 27th annual international conference of the ieee engineering in medicine and biology society 2005 (IEEE EMBC 2005). Shanghai, China. IEEE Press.

[17] 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.

[18] 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.

[19] 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.

[20] 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.