Deployment of Protein Ontology Framework

Amandeep Sidhu S[1], Tharam Dillon S[2] and Elizabeth Chang[3]
[1,2,3]Curtin University of Technology, Perth, Australia
(Amandeep.Sidhu, Tharam.Dillon, Elizabeth.Chang)@cbs.curtin.edu.au

Abstract

Bioinformatics researchers have long identified the need for interoperation among protein databases, knowledge bases and other information sources. Despite advances, interoperation among knowledge and data sources is still an issue with every new protein data and information source that is created. We proposed Protein Ontology (PO) in 2003 (see: http://www.proteinontology.info/) that provides an efficient integration and interoperation framework among existing protein data and information sources. In this paper we explore its design, current issue its development, and its growing adoption as a standard for representation in biomedical domain.

Key words: Prion Proteins, PDB Format, Algorithms.

I. INTRODUCTION

Biology demonstrates three challenges for data integration that are common in evolving scientific domains but not typically found elsewhere. The first is the sheer number of available data sources and the inherent heterogeneity of their contents. Some of these sources contain data from a single lab or project, whereas others are the definitive repositories for very specific types of information (e.g., for a specific genetic mutation). Not only do these sources complicate the concept identification issue previously mentioned (because they use highly specialized data semantics), but their make it infeasible to incorporate all of them into a consistent repository. Second, the data formats and data access methods change regularly. These changes are an attempt to keep up with the scientific evolution occurring in the community at large. However, a change in a data source representation can have dramatic impact on systems that integrate that source, causing the integration to fail on the new format. Third, the data and related analysis are becoming increasingly complex. As the nature of genomics and proteomics research evolves from a predominantly wet-lab activity into knowledge-based analysis, the scientists' need to access the wide variety of available information increases dramatically. To address this need, information needs to be brought together from various heterogeneous data sources and presented to researchers in ways that allow them to answer their questions.

Problems facing genomics and proteomics data are related to data semantics the meaning of data represented in a data source and the difference between semantics within a set of sources [1]. Unfortunately, the semantics of biological data are usually hard to define precisely because they are not explicitly stated but are implicitly included in the database design. The reason is simple: At a given time, within a single research community, common definitions of various terms are often well understood and have precise meaning. As a result, those within that community usually understand the semantics of a data source without needing to be explicitly defined. However, proteomics (much less all of biology or life science) is not a single, consistent scientific domain; it is composed of dozens of smaller, focused research communities [2]. This would not be a significant issue if researchers only accessed data from within a single domain, but that is not usually the case. Typically, researchers require integrated access to data from multiple domains, which requires resolving terms that have slightly different meanings across the communities. This is further complicated by the observations that the specific community whose terminology is being used by the data source is usually not explicitly identified and that the terminology evolves over time [3].

For proteomics, domain users frequently use web sites as sources of protein data, but often fail to retrieve the correct information due to the heterogeneous and complex structure of the data formats. Recent progress in proteomics, computational biology, and ontology development has presented an opportunity to investigate protein data sources from a unique perspective that is, examining protein data sources through structure and hierarchy of Protein Ontology (PO) [12-21]. Various data mining algorithms and mathematical models provide methods for analyzing protein data sources; however, there are two issues that need to be addressed: (1) the need for standards for defining protein data description

and exchange and (2) eliminating errors which arise with the data integration methodologies for complex queries. Protein Ontology is designed to meet these needs by providing a structured protein data specification for Protein Data Representation. Protein Ontology is a standard for representing protein data in a way that helps in defining data integration and data mining models for Protein Structure and Function. Protein Ontology provides a vocabulary for representing knowledge about the proteomics domain and describes specific data sources therein. The role of protein ontology is to create explicit specified conceptualizations that can be shared, reused, and integrated in the analysis and design stages of information and knowledge systems for bioinformatics.

Biomedical Knowledge of Proteomics Domain is represented in the Protein Ontology, whose instantiations, which are undergoing evolution, need a good management and maintenance system. Protein Ontology instantiations signify data and information about proteins that is shared and has evolved to reflect development in the Protein Ontology Project and the Proteomics Domain itself. Protein Data and Knowledge captured in Protein Ontology Concepts and Instantiations represents abstraction of data sources and expertise in the proteomics domain. Abstraction is divided into generic and derived concepts of protein ontology. Protein Ontology instantiations are derived as a result of populating protein data and information and are referred to as instances of protein ontology classes. Instantiations are also known as instance knowledge of the protein ontology. The instantiations of PO represent knowledge about respective proteins. Concrete data instances about various proteins from underlying diverse protein data and knowledge sources are stored as PO instantiations in the PO Instance Store.

The Protein Ontology Instance Store is created as a repository for existing protein data using the PO format. PO uses data sources that include new proteome information resources like PDB [4], SCOP [7], and RESID [8] as well as traditional sources of information where information is maintained in a knowledge base of scientific text files like OMIM [10] and from various published scientific literature in various journals. The PO Instance Store is represented using OWL. All the Protein Ontology Instances are available for download (http://proteinontology.info/proteins.htm) in OWL [22] format that can be read by any popular editor like Protégé (http://protege.stanford.edu/).

II. PO INSTANTIATIONS TRANSFORMATION

In this section, we report in particular on how protein data are transformed or mapped into concepts formed in the protein ontology as instance knowledge. Protein Ontology Web Retrieval System (PO-WEB) manages the connection between Protein Ontology Conceptual Framework [19] and the Protein Ontology Instance Store. PO-WEB is built on top of Jena [9], which we would like to gratefully acknowledge. Jena, developed by the Hewlett-Packard Company, is a Java framework with the capacity to manipulate ontologies. The version of Jena used is Jena 2.1. PO-WEB provides acquisition, navigation, and querying of the Protein Ontology Instance Store.

A. Acquisition

In many cases, creators use different data descriptors to refer to same real-world protein data. For example creators of PDB [4, 5] use the terms *organ, tissue, and organelle* to specify the location of a protein molecule; whereas creators of SWISS-PROT [6] use the terms *subcellular location and tissue specificity* for the same. Without knowing that these terms mean similar things, a researcher will miss important information about the protein under study.

The process of acquiring data and knowledge from the proteomics domain is described in this stage, which applies algorithms and methods analyzing protein data files and proteomics domain texts. The terminology used by domain experts is defined in protein ontology. In this study, in order to collect a glossary of concepts (classes) for the proteomics domain, firstly, an analysis was performed on 4 major protein data sources: PDB [5], SWISS-PROT [6], SCOP [7], and PIR [11]. For example, Atom records in Protein Data Bank in PDB format [5] present the atomic coordinates for standard residues. They also present the occupancy and temperature factor for each atom. A typical Atom Record is shown in Fig. 1.

```
ATOM      1  N   LEU L   2      84.209  53.240  23.108  1.00 90.00           N
ATOM      2  CA  LEU L   2      83.250  52.895  24.194  1.00 90.00           C
ATOM      3  C   LEU L   2      82.139  52.003  23.650  1.00 90.00           C
ATOM      4  O   LEU L   2      81.406  52.395  22.734  1.00 90.00           O
ATOM      5  CB  LEU L   2      82.448  54.156  24.834  1.00 78.32           C
ATOM      6  CG  LEU L   2      81.639  53.954  25.976  1.00 78.32           C
ATOM      7  CD1 LEU L   2      82.216  53.079  27.079  1.00 78.32           C
ATOM      8  CD2 LEU L   2      81.241  55.292  26.530  1.00 78.32           C
ATOM      9  N   VAL L   3      82.008  50.815  24.230  1.00 90.00           N
ATOM     10  CA  VAL L   3      80.997  49.853  23.804  1.00 90.00           C
ATOM     11  C   VAL L   3      80.133  49.307  24.935  1.00 90.00           C
ATOM     12  O   VAL L   3      80.611  48.572  25.798  1.00 90.00           O
ATOM     13  CB  VAL L   3      81.642  48.637  23.094  1.00 86.42           C
ATOM     14  CG1 VAL L   3      80.564  47.766  22.462  1.00 86.42           C
ATOM     15  CG2 VAL L   3      82.651  49.089  22.049  1.00 86.42           C
ATOM     16  N   MET L   4      78.852  49.656  24.908  1.00 65.25           N
ATOM     17  CA  MET L   4      77.902  49.170  25.897  1.00 65.25           C
ATOM     18  C   MET L   4      77.502  47.750  25.477  1.00 65.25           C
ATOM     19  O   MET L   4      76.977  47.545  24.381  1.00 65.25           O
```

Fig.1 Protein Atoms described in PDB format

An interface (Fig.2) is used to parse the data from various protein data sources like PDB and unify them in the PO format. Protein data is parsed according OWL schema specifications.

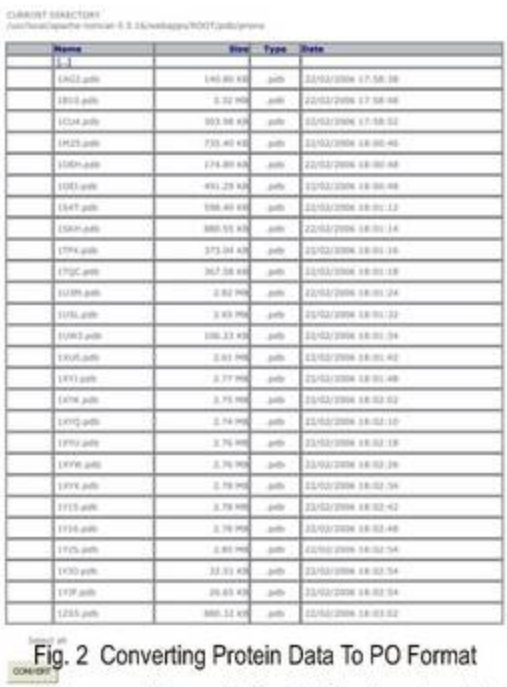


Fig. 2 Converting Protein Data To PO Format

In this case, Atoms of a Protein Structure described in PDB format are converted using this interface to an instance of *Atom Concept* stored in PO Instance Store (Figure 3) represented using OWL Web Ontology Language.

Fig. 3: Atom Concept from PO Instance Store

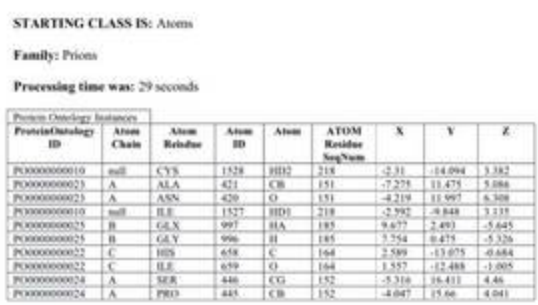
B. Navigation

In this section, we deal with the accessing of knowledge held in the PO Instance Store. Protein Ontology concept structures are formulated so that they can easily be navigated. The knowledge is provided in hierarchical form so upper level concepts or lower level concepts or adjacent concepts can easily be navigated. Technically for this function, PO-WEB focuses on the Protein Ontology Schema in OWL, and the set of statements that comprises the abstraction and instantiations. To navigate the PO Instance Store, PO-WEB reads Protein Ontology Schema in OWL and then accesses the individual instances of the elements. PO Conceptual Hierarchy with their brief description is at: http://proteinontology.info/hierarchy.htm

C. Querying

Protein Ontology can be formally represented using an ontology language such as Ontology Web Language (OWL) or Resource Description Format (RDF). Querying Protein Ontology involves the formalization of each term and the constraints used by the ontology. Some of the formalisms not provided by OWL in which Protein Ontology is represented are defined using Protein Ontology Algebra [18]. Terms are represented through classes, relations, functions, and instances. Queries to extract Protein Ontology Concepts and Instances are also formulated using Protein Ontology Specification and Algebra (Fig. 4 and 5).

Fig. 4 Concept and Attribute Selection



ProteinOntology ID	Atom Chain	Atom Residue	Atom ID	Atom	ATOM Residue SeqNum	X	Y	Z
PO0000000010	null	CYS	1528	HD2	218	-2.31	-14.094	3.382
PO0000000023	A	ALA	421	CB	151	-7.275	11.475	5.086
PO0000000023	A	ASN	420	O	151	-4.219	11.997	6.308
PO0000000010	null	ILE	1527	HD1	218	-2.592	-9.848	3.135
PO0000000025	B	GLX	997	HA	185	9.677	2.493	-5.645
PO0000000025	B	GLY	996	H	185	7.754	0.475	-5.326
PO0000000022	C	HIS	658	C	164	2.589	-13.075	-0.684
PO0000000022	C	ILE	659	O	164	1.557	-12.488	-1.005
PO0000000024	A	SER	446	CG	152	-5.316	16.411	4.46
PO0000000024	A	PRO	445	CB	152	-4.047	15.66	4.041

Fig. 5 Query Results for the Selection

III. CASE STUDY: PRION PROTEINS

Prion - short for proteinaceous infectious particle (-on) that lacks nucleic acid (by analogy to virion) - is a type of infectious agent composed only of protein. It causes a number of diseases in a variety of animals, including bovine spongiform encephalopathy (BSE, also known as mad cow disease) in cattle and Creutzfeldt-Jakob

disease in humans. All Prion diseases affect the structure of the brain or other neural tissue, and all are untreatable and fatal. All major Prion Proteins are available for download from the PO Instance Store (http://proteinontology.info/proteins.htm).

A. Illustrative Example

In this section, examine how information is extracted from PDB [ref] for a Prion Protein using Protein Ontology. In this example, we show the conversion of Entry and Three-dimensional atomic structures, and like B. Subtilis discussed earlier, conversion of all data from PDB to Protein Ontology is carried out. In the next section, we will run tree mining algorithms on the three dimensional protein structure data of Prion Proteins stored in the PO Instance Store to study the similarity of patterns among proteins.

Basic Information about Prion Protein, which is described in terms of Title, Keywords, Experimental Method (EXPDTA), Source and Authors is fetched from Protein Data Bank (PDB ID: 1SKH). This information is represented in PDB format as follows:

```
HEADER    UNKNOWN FUNCTION                        05-MAR-04   1SKH
TITLE     N-TERMINAL (1-30) OF BOVINE PRION PROTEIN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MAJOR PRION PROTEIN 2;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: N-TERMINAL DOMAIN (RESIDUES 1-30);
COMPND   5 SYNONYM: PRP, MAJOR SCRAPIE-ASSOCIATED FIBRIL PROTEIN 2
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: BOS TAURUS;
SOURCE   3 ORGANISM_COMMON: BOVINE
KEYWDS    COIL-HELIX-COIL
EXPDTA    NMR, 22 STRUCTURES
AUTHOR    H.BIVERSTAHL,A.ANDERSSON,A.GRASLUND,L.MALER
```

Fig. 6 Protein Entry Information in PDB

This information is extracted from PDB and described in Description Concept of Protein Ontology in a Web Ontology Language representation as follows:

```
<Description rdf:about="http://www.proteinontology.info/po.owl#DescriptionInstance">
  <Authors rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >H.BIVERSTAHL,A.ANDERSSON,A.GRASLUND,L.MALER</Authors>
  <Classification rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >UNKNOWN FUNCTION</Classification>
  <Title rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >N-TERMINAL (1-30) OF BOVINE PRION PROTEIN</Title>
  <SourceDatabaseID rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >1SKH</SourceDatabaseID>
  <Experiment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >NMR, 22 STRUCTURES</Experiment>
  <SourceDatabaseName rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Protein Data Bank - PDB</SourceDatabaseName>
  <Keywords rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >COIL-HELIX-COIL</Keywords>
  <SourceSubmissionDate rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >05-MAR-04</SourceSubmissionDate>
</Description>
```

Fig. 7 Description Concept in PO

Organism and Cellular Source where protein resides from PDB is described using Source Cell Concept of Protein Ontology, and represented in Web ontology Language as follows:

```
<SourceCell rdf:about="http://www.proteinontology.info/po.owl#SourceCellInstance46">
  <OrganismCommon rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >: BOVINE</OrganismCommon>
  <SourceMoleculeID rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >1</SourceMoleculeID>
</SourceCell>
```

Fig. 8 Source Cell Concept in PO

Lastly, Atomic Coordinates for Standard Residues in a Protein Structure are represented by Atom Records in PDB as follows:

```
ATOM      3  C   MET A   1     -18.582  -1.859  -2.798  1.00  0.00           C
ATOM      4  O   MET A   1     -19.565  -2.236  -3.281  1.00  0.00           O
ATOM      5  CB  MET A   1     -16.567  -2.815  -1.532  1.00  0.00           C
ATOM      6  CG  MET A   1     -15.738  -4.095  -1.414  1.00  0.00           C
```

Fig. 9 Atom Records in PDB

The Atom Record (Atom ID: 3) is described in Protein Ontology as AtomInstance297170 and is represented in Web Ontology Language as follows:

```
<Atom rdf:about="http://www.owl-ontologies.com/ProteinOntology.owl#AtomInstance297170">
  <Occupancy rdf:datatype="http://www.w3.org/2001/XMLSchema#float">  1.00</Occupancy>
  <Y rdf:datatype="http://www.w3.org/2001/XMLSchema#float">  -1.859</Y>
  <Atom rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">    3</Atom>
  <AtomName rdf:datatype="http://www.w3.org/2001/XMLSchema#string"> C   </AtomName>
   <Element rdf:datatype="http://www.w3.org/2001/XMLSchema#string" >C</Element>
   <TemperatureFactor rdf:datatype="http://www.w3.org/2001/XMLSchema#float" >  0.00</
TemperatureFactor>
  <X rdf:datatype="http://www.w3.org/2001/XMLSchema#float"> -18.582</X>
   <AtomResSeqNum rdf:datatype="http://www.w3.org/2001/XMLSchema#int">   1</AtomResSeqNum>
   <Z rdf:datatype="http://www.w3.org/2001/XMLSchema#float">  -2.798</Z>
   <AtomResidueRef>
  <Residue rdf:ID="MET">
  <Residue rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >MET</Residue>
  <ResidueName rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Methionine</ResidueName>
  </Residue>
  </AtomResidueRef>
  <AtomChainRef>
  <Chains rdf:ID="A">
  <ChainName rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Chain A</ChainName>
  <Chain rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >A</Chain>
  </Chains>
  </AtomChainRef>
  </Atom>
```

Fig. 10 Atoms Concept in PO

B. Tree Mining PO Instances of Prion Proteins

Tree Mining has attracted much interest among the data mining community, due to the increasing use of semi-structured data sources for more meaningful knowledge representations. This is particularly evident in areas such as Bioinformatics, XML Mining, Web applications, scientific data management, and more generally in any area where the knowledge is represented in a tree-structured form. Our group's work in the area of frequent subtree mining is characterized by adopting a Tree Model Guided (TMG) candidate generation [37, 39] as opposed to the join approach, which is commonly used. This non-redundant systematic enumeration model ensures that only valid candidates are generated which conform to the actual tree structure of the data. Furthermore, our unique Embedding List representation of the tree structure has allowed for an efficient implementation of the TMG approach which has resulted in efficient algorithms for mining embedded subtrees (MB3) [37] and induced subtrees (IMB3) [38], from a databases of labeled rooted

ordered subtrees. MB3-R and IMB3-R algorithms [39] are the latest implementations that adopt a more space efficient global representation and store only the right most path information for candidate subtrees.

e apply the MB3-R algorithm to the Prions dataset [36] in order to extract the frequently occurring subtrees and to check for efficiency of data mining algorithms on PO Instance Store. Prions dataset describes PO Instance Store for Human Prion proteins in XML format [12, 13]. The XML tags are first mapped to integer indexes similar to the format used in TMGJ [37]. Representing a label as an integer instead of a string label has considerable performance and space-saving advantages. Since the maximum height of the Prions tree structure is 1, all candidate sub trees generated are induced sub trees. The experiments were run on 3Ghz (Intel-CPU), 2Gb RAM, Mandrake 10.2 Linux machine and compilation was performed using GNU g++ (3.4.3) with g and O3 parameters. Occurrence-match support definition was used. The total run-time and memory usage of the MB3 algorithm is displayed in Fig. 11, for varying support thresholds.

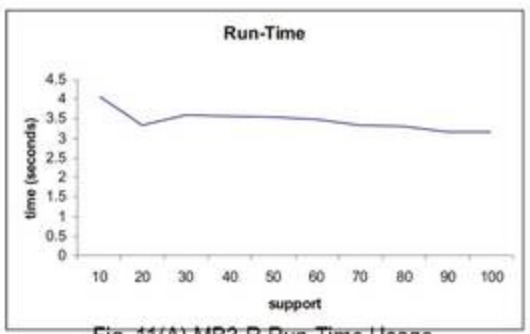


Fig. 11(A) MB3-R Run-Time Usage

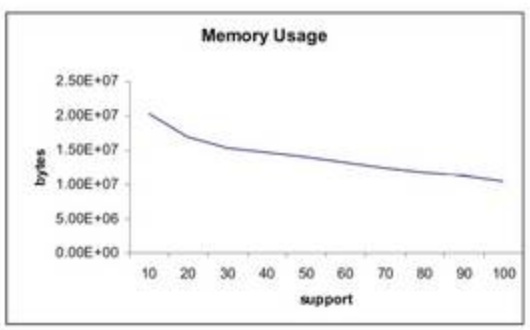


Fig. 11(B) MB3-R Memory Usage

We also used some standard hierarchical and tree mining algorithms (Tan et al., 2006a) on the PO instance store. We compared our MB3-Miner (MB3) algorithm with X3-Miner (X3), VTreeMiner (VTM) and PatternMatcher (PM) for mining embedded subtrees and our IMB3-Miner (IMB3) with FREQT (FT) for mining induced subtrees of PO instance store. Figure 12 shows the time performance of different algorithms.

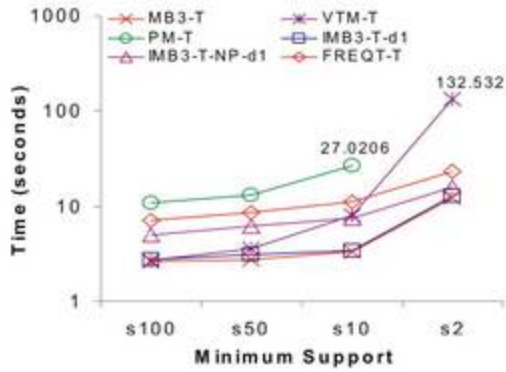


Fig. 12 Performance of Tree Mining Algorithms

Quite interestingly, with the Prion dataset of PO, the number of frequent candidate sub trees generated is identical for all algorithms (Fig. 13).

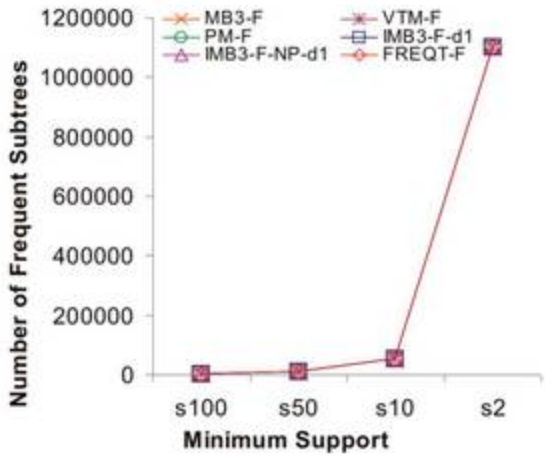


Fig.13 Frequent Sub Trees for Prion Dataset

IV. ADOPTION OF PROTEIN ONTOLOGY AS A STANDARD

Protein Ontology is a part of Standardized Biomedical Ontologies available through the National Center for Biomedical Ontologies [23] along with Gene Ontology [24], Flybase [25], and others (http://cbioapprd.stanford.edu/ncbo/ faces/pages/ontology_list.xhtml). Also different research groups are using Protein Ontology for different purposes. Y. Wang et al. [26] shows Protein Ontology as an example of a structured approach for knowledge modeling providing solid inference and retrieval functionalities. F. Porto [27] discusses Protein Ontology in his report under Ontologies for Bioinformatics. H. Tan et al. [36] use Protein Ontology generated data set to evaluate their algorithms.

A. Kupfer [28] use Protein Ontology along with Gene Ontology to understand concepts when discussing a coevolution approach for database schemas. N. Bolshakova et al. [29] discuss protein ontology under a section on Biomedical Ontologies while comparing data based and ontology based approaches for cluster validation of microarrays. IQlue [30] references Protein Ontology when reviewing Ontology Development in their white paper. L. Dhanapalan and Jake Y. Chen [31] discuss protein ontology in detail when doing case study of integrating protein interaction data using semantic web technology. K. Pinagé and V. Brilhante [32] used Protein Ontology for Protein Structure Homology Modeling. Just Recently researchers [33] discuss in detail Protein Ontology along with other major biomedical ontologies, while proposing a text mining based ontology construction methodology for Protein Data mainly for PIR database. A. Kupfer et al. [34] reuses the concept of chains from Protein Ontology when proposing database ontology for signal transduction pathways. Lastly, Z. Lacroix et al. [35] discuss Protein Ontology briefly when proposing a semantic model to integrate biological resources.

V. FUTURE WORK

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. Also 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.

In reality, protein data sources are updated over a period of time to reflect development in proteomics. Since changes are inevitable during proteomics experimentations, the Protein Ontology Instance Store is constantly confronted with the evolution problem. If such changes are not properly traced or maintained, this would impede the use of the protein ontology. Therefore a semi-automatic process becomes increasingly necessary to facilitate updating tasks and to ensure reliability. The evolution problem of the PO Instance Store can be handled partly by using the Difference Operator of PO Algebra [18]. It will suggest whether instances have not been entered properly or whether there has been any change to the underlying protein data and knowledge sources from which the PO Instance Store is populated.

VI. SUMMARY

In this paper, we discuss the implementation methodology for the Protein Ontology based on information systems research methodology. Here, we explore the process of instantiations transformation from Protein Data Sources to Protein Ontology Instance Store. As a proof of concept, we also discuss case study of Prions protein family. We also show the accuracy and performance of the popular data mining algorithms on Prion Dataset.

This paper also outlines various strengths of the Protein Ontology that make it an extremely useful tool for protein data integration and data mining. The Protein Ontology Project has been cited in over 60 scientific publications to date. Here, we also discuss the prominent usage of the Protein Ontology by some of research groups. Lastly, we discuss in detail the future challenges that are being addressed by the Protein Ontology Project. The Protein Ontology is the first ontology of its kind that was proposed for the purpose of integrating protein data and information sources on this scale and has tremendous potential to address challenges of heterogeneity, data integration and interoperability specifically in proteomics and generally in the biomedical domain.

REFERENCES

[1] S. E. Brenner, "World Wide Web and molecular biology," *Science*, vol. 268, pp. 622-623, 1995

[2] A. Baxevanis, "The Molecular Biology Data Collection: 2002 update," *Nucleic Acids Research*, vol. 30, pp. 1-12, 2002.

[3] P. Buneman, S. Davidson, K. Hart, C. Overton, and L. Wong, "A Data Transformation System for Biological Data Sources.," presented at 21st International Conference on Very Large Data Bases (VLDB 1995), Zurich, Switzerland, 1995.

[4] J. Wesbrook, Z. Feng, S. Jain, T. N. Bhat, N. Thanki, V. Ravichandran, G. L. Gilliland, W. F. Bluhm, H. Weissig, D. S. Greer, P. E. Bourne, and H. M. Berman, "The Protein Data Bank: unifying the archive," *Nucleic Acids Research*, vol. 30, pp. 245-248, 2002.

[5] J. Westbrook and P. M. D. Fitzgerald, "The PDB format, mmCIF formats and other data formats," in *Structural Bioinformatics*, P. E. Bourne and H. Weissig, Eds. Hoboken, NJ: John Wiley & Sons, Inc., 2003, pp. 161-179.

[6] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its supplement

TrEMBL," *Nucleic Acids Research*, vol. 25, pp. 31-36, 1997.

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," Journal of Molecular Biology, vol. 247, pp. 536-540, 1995.

[8] J. S. Garavelli, "The RESID Database of Protein Modifications: 2003 developments," Nucleic Acids Research, vol. 31, pp. 499-501, 2003.

[9] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson, "Jena: implementing the semantic web recommendations," Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, pp. 74-83, 2004.

[10] V. A. McKusick, *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*. Baltimore: Johns Hopkins University Press, 1998.

[11] W. C. Barker, J. S. Garavelli, D. H. Haft, L. T. Hunt, C. R. Marzec, and B. C. Orcutt, "The PIR-International Protein Sequence Database," Nucleic Acids Research, vol. 26, pp. 27-32, 1998.

[12] A. S. Sidhu, T. S. Dillon, H. Setiawan, and B. S. Sidhu, "Comprehensive Protein Database Representation," presented at 8th International Conference on Research in Computational Molecular Biology 2004 (RECOMB 2004), San Diego, California, 2004.

[13] A. S. Sidhu, T. S. Dillon, B. S. Sidhu, and H. Setiawan, "A Unified Representation of Protein Structure Databases," in *Biotechnological Approaches for Sustainable Development*, M. S. Reddy and S. Khanna, Eds. India: Allied Publishers, 2004, pp. 396-408.

[14] A. S. Sidhu, T. S. Dillon, and E. Chang, "Ontological Foundation for Protein Data Models," presented at 1st IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2005), In conjunction with On The Move Federated Conferences (OTM 2005), Agia Napa, Cyprus, 2005.

[15] A. S. Sidhu, T. S. Dillon, and E. Chang, "An Ontology for Protein Data Models," presented at 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2005 (IEEE EMBC 2005), Shanghai, China, 2005.

[16] A. S. Sidhu, T. S. Dillon, and E. Chang, "Advances in Protein Ontology Project," presented at 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, 2006.

[17] A. S. Sidhu, T. S. Dillon, and E. Chang, "Integration of Protein Data Sources through PO," presented at 17th International Conference on Database and Expert Systems Applications (DEXA 2006), Poland, 2006.

[18] A. S. Sidhu, T. S. Dillon, and E. Chang, "Towards Semantic Interoperability of Protein Data Sources," presented at 2nd IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2006) in conjuction with OTM 2006, France, 2006.

[19] A. S. Sidhu, T. S. Dillon, and E. Chang, "Protei Ontology," in *Biological Database Modeling*, J. Chen and A. S. Sidhu, Eds. New York: Artech House, 2007, pp. 39-60.

[20] A. S. Sidhu, T. S. Dillon, E. Chang, and B. S. Sidhu, "Protein ontology: vocabulary for protein data," presented at 3rd International IEEE Conference on Information Technology and Applications, 2005 (IEEE ICITA 2005), Sydney, 2005.

[21] A. S. Sidhu, T. S. Dillon, B. S. Sidhu, and H. Setiawan, "An XML based semantic protein map," presented at 5th International Conference on Data Mining, Text Mining and their Business Applications (Data Mining 2004), Malaga, Spain, 2004.

[22] W3C-OWLGuide, "OWL Web Ontology Language Guide," in *W3C Recommendation 10 February 2004*, M. K. Smith, C. Welty, and D. L. McGuinness, Eds.: World Wide Web Consortium, 2004.

[23] D. L. Rubin, S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C. G. Ghute, H. Solbrig, M. Storey, B. Smith, J. Day-Richter, N. F. Noy, and M. A. Musen, "National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge," *OMICS A Journal of Integrative Biology*, vol. 10, pp. 185-198, 2006.

[24] M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, J. C. Cherry, J. Corradi, and K. Dolinski, "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, vol. 11, pp. 1425-1433, 2001.

25] M. Ashburner, "FlyBase," *Genome News*, vol. 13, pp. 19-20, 1993.

[26] Y. Wang, J. Wang, and S. Zhang, "Collaborative knowledge management by integrating knowledge modeling and workflow modeling," presented at IEEE International Conference on Information Reuse and Integration (IRI 2005), Las Vegas, Nevada, USA, 2005.

[27] F. Porto, "Reasoning on Dynamically Built Reasoning Space with Ontology Modules," *LECTURE NOTES IN COMPUTER SCIENCE*, vol. 3761, pp. 1623-1638, 2005.

[28] A. Kupfer, S. Eckstein, K. Neumann, and B. Mathiak, "A Coevolution Approach for Database Schemas and Related Ontologies," presented at 19th IEEE Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, 2006.

[29] N. Bolshakova, A. Zamolotskikh, and P. Cunningham, "Comparison of the Data-based and Gene Ontology-based Approaches to Cluster Validation Methods for Gene Microarrays," presented at 19th IEEE Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, 2006.

[30] IQlue, "ONTOLOGY: "The specification of a shared conceptualization" - A Review Document," IQlue, a division of siOnet Ltd, Herzelia, Israel 2006.

[31] L. Dhanapalan and J. Y. Chen, "A Case Study of Integrating Protein Interaction Data using Semantic Web Technology," *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, vol. 3, 2007.

[32] K. Pinagé and V. Brilhante, "Protein Structure Homology Modelling assisted by Ontology," presented at 14th Annual International conference on Intelligent Systems for Molecular Biology (ISMB 2006), Fortaleza, Brazil, 2006.

[33] D. A. Natale, C. N. Arighi, W. Barker, J. Blake, T. Chang, Z. Hu, H. Liu, B. Smith, and C. H. Wu, "Framework for a Protein Ontology," presented at ACM First International Workshop on Text Mining in Bioinformatics (TMBIO 2006), Arlington, Virginia, 2006.

[34] A. Kupfer, S. Eckstein, B. Stormann, and B. Mathiak, "A database ontology for signal transduction pathways," *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, vol. 3, 2007.

[35] Z. Lacroix, L. Raschid, and M. E. Vidal, "Semantic Model to Integrate Biological Resources," presented at 3rd Semantic Web and Databases Workshop with ICDE 2006, Atlanta, USA, 2006.

[36] F. Hadzic, T. S. Dillon, A. S. Sidhu, E. Chang, and H. Tan, "Mining Substructures in Protein Data," in 2006 IEEE Workshop on Data Mining in Bioinformatics (DMB 2006) in conjunction with 6th IEEE ICDM 2006, Hong Kong, 2006.

[37] H. Tan, T. S. Dillon, F. Hadzic, E. Chang, and L. Feng, "MB3 Miner: mining eMBedded sub-TREEs using Tree Model Guided candidate generation," in 1st International Workshop on Mining Complex Data, held in conjunction with ICDM 2005, Texas, USA, 2005.

[38] H. Tan, T. S. Dillon, F. Hadzic, E. Chang, and L. Feng, "IMB3-Miner: Mining Induced/Embedded Subtrees by Constraining the Level of Embedding," in 10th Pacific-Asia Knowledge Discovery and Data Mining Conference (PAKDD 2006), Singapore, 2006, pp. 450-461.

[39] H. Tan, T. S. Dillon, F. Hadzic, E. Chang, and L. Feng, "Mining induced/embedded subtrees using the level of embedding constraint," Knowledge and Information Systems An International Journal, 2006.

[40] H. Tan, T. S. Dillon, F. Hadzic, L. Feng, and E. Chang, "Tree Model Guided Candidate Generation for Mining Frequent Subtrees from XML," ACM Transactions on Knowledge Discovery from Data (TKDD), 2006