Trust based Decision Making Approach for Protein Ontology

Amandeep S. Sidhu, Member, IEEE, Farookh K. Hussain, Tharam S. Dillon, Fellow, IEEE, Elizabeth Chang, Senior Member, IEEE

Digital Ecosystems and Business Intelligence Institute,
Curtin University of Technology Perth, Australia
(Amandeep.Sidhu, Farookh.Hussain, Tharam.Dillon, Elizabeth.Chang)@cbs.curtin.edu.au

Abstract

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 information about proteins that is shared and has evolved to reflect development in Protein Ontology Project and Proteomics Domain itself. In this paper we explore the development of a conceptual framework for Protein Ontology instantiations management by using the concepts of trust and reputation in Biomedical Domain. The developed and engineered ontology approach is trustworthy and facilitates reliable additions and updates to the Protein Ontology.

1. Introduction

The term 'ontology' has its origins in metaphysics and philosophical sciences. Ontology refers to the old philosophical discipline introduced by Aristotle [1]. Ontology as a branch of philosophy is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality. The concept of ontology was first borrowed from the realm of Philosophy by Artificial Intelligence researchers and has since become a matter of interest to computer and information scientists in general. In computer science literature, the term takes on a new meaning, but one that is not entirely unrelated to its philosophical counterpart. There are many different ontology definitions in the computer and information science literature [2]. But, all researchers agree on the importance of ontology research in terms of the necessary mechanisms to represent, share and reuse the existing domain knowledge [3]. Ontologies in biomedicine [4, 5] have emerged because of the need for common shared vocabularies for effective communication across diverse sources of biological data and knowledge. These shared vocabularies usually include concepts, relationships between concepts, definitions of these concepts and relationships and also the possibility of defining ontology rules and axioms, in order to define a mechanism to control the objects that can be introduced in the ontology and to apply logical inference.

We proposed Protein Ontology [6-10] or PO for short in 2003 for Proteomics Domain that 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.

Protein Data and Knowledge captured in Protein Ontology Concepts and Instantiations, represents abstraction of data sources and expertise in proteomics domain. Abstraction is divided into generic and derived concepts of 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 PO Instance Store.

2. Protein Ontology Basics

In order to understand development of Trustworthy PO for the context of this paper in this section we discuss a

1-4244-0826-1/07/$20.00 © 2007 IEEE

brief overview of PO Conceptual Framework Protein Ontology Instance Store.

2.1 Protein Ontology Conceptual Framework

The ultimate goal of Protein Ontology (PO) is to deduce from proteomics data all its biological features and describing all intermediate structures: primary amino acid sequence, secondary structure folds and domains, tertiary three dimensional atomic structure, quaternary active functional sites, etc. Thus, complete protein annotation for all types of proteins for an organism is a very complex process that requires besides extracting data from various protein databases, integration of additional information: results of protein experiments, analysis of bioinformatics tools, and biological knowledge accumulated over years. This constitutes a huge mass of heterogeneous protein data sources that need to rightly represented and stored. Protein Annotators must be able to readily retrieve and consult these data. Therefore protein databases and man-machine interfaces are very important when defining a protein annotation using protein ontology.

The process of development of a protein annotation based on our protein ontology requires an important effort to organize, standardize and rationalize protein data and concepts. First of all, protein information must be defined and organized in a systematic manner in databases. In this context, PO addresses the following problems of existing protein databases: redundancy, data quality (errors, incorrect annotations, and inconsistencies), lack of standardization in nomenclature etc. The process of annotation relies heavily on integration of heterogeneous protein data. Integration is thus a key concept if one wants to make full use of protein data from collections. In order to be able to integrate various protein data it is important that concepts underlying the data be agreed upon by community. PO provides a framework of structured vocabularies and standardized description of protein concepts that helps to achieve this agreement and achieve uniformity in protein data representation.

Protein Ontology or PO consists of concepts (or classes), which are data descriptors for proteomics data and the relationships 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; (3) a set of relationships between classes to link concepts in ontology in more complicated ways then implied by the hierarchy, to promote reuse of concepts in the ontology; and (4) a set of algebraic operators for querying protein ontology instances. An overview of concepts of PO is shown in the following diagram.

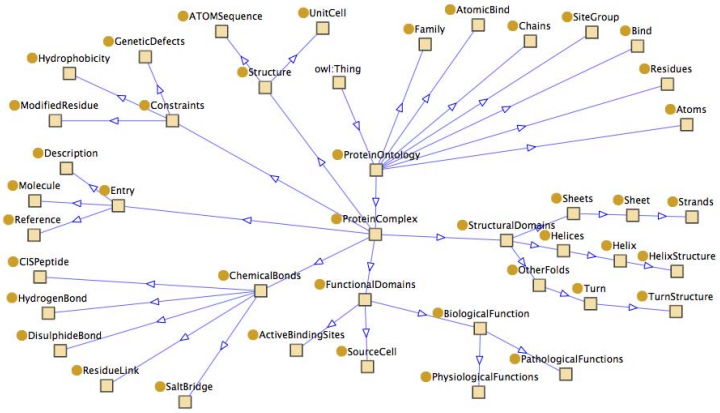

Figure 1: Protein Ontology Concepts

PO provides technical and scientific infrastructure to allow evidence based description and analysis of relationships between proteins. More details about PO Development can be found in [8] and on the website: http://www.proteinontology.info/.

2.2 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, SCOP, and RESID as well as traditional sources of information where information is maintained in a knowledge base of scientific text files like OMIM and from various published scientific literature in various journals. The PO Instance Store is represented using Web Ontology Language. Currently, the PO Instance Store contains data instances of following protein families for B.Subtilis and Prion Proteins. More protein data instances will be added as the PO is further developed. All the Protein Ontology Instances are available for download (http://proteinontology.info/proteins.htm) in OWL format that can be read by any popular editor like Protégé (http://protege.stanford.edu/).

3. Engineering Trustworthy PO

Here we discuss a conceptual framework that we are working on, to engineer Trustworthy Protein Ontology. It is termed as 'Trustworthy Protein Ontology' as the final engineered ontology is trustworthy in the sense that it is accurate and precise. The final engineered ontology does not contain any redundant, inconsistent, and incorrect data or relationships.

Consider the scenario where we have 'N' Researchers. Each of these Researchers enters the data into an Intermediate Protein Ontology (IPO). IPO is

mirror of the Original PO and contains same concepts in an exactly similar structured hierarchy as PO. However the research assistants may not be necessarily the experts in field of proteomics for which the ontology is being engineered. Hence we propose that instead of allowing research assistants to make changes directly to the Original PO, changes should be entered into the IPO. PO administrator then goes through IPO to check if the concepts, relationships and instances entered by research assistants. PO administrator is a person who is an expert in the field of proteomics for which trustworthy PO is engineered. PO administrator has knowledge about data formats of diverse protein data and knowledge sources. After research assistants enter the data in IPO, PO administrator goes through IPO in order skim out concepts, relationships and instances which are redundant, inconsistent, and incorrect. This is done by running syntax and semantic checks on IPO, to check its validity in regards to concepts, relationships and instances already present in Original PO. There are two ways in which PO administrator may choose to skim through IPO.

Method 1: PO administrator goes through the whole IPO to which changes have been submitted by the Research Assistants to determine those concepts, relationships and instances which are redundant, inconsistent, and incorrect. PO administrator then removes or fixes these concepts, relationships and instances to create the final engineered IPO. Once all discrepancies have been removed from the final engineered IPO, and it has been checked for validity with the Original PO, all the changes made to IPO are integrated into the Original PO. This method compares structure and relationships of IPO and Original PO. This method is tedious and requires a lot of time and effort by the PO administrator. PO administrators can alternatively choose Method 2 as a means to engineer trustworthy ontology which is quick, effective and does all the checks.

Method 2: PO administrator uses an administration console to skim through IPO using a defined set of rules that denotes what a correct concept would be, what a correct relationship between those concepts would be and what a correct instance of the concept would be. These set of rules utilize structure and semantics of PO to facilitate validation of any changes made to IPO by research assistants. PO structured vocabulary briefly outlined in Section 2 has 92 pre-defined concepts that belong to set of valid concepts, **SET V**. Of these 92 concepts, 12 concepts are necessary to define the basic information to enter protein complex data into the PO framework. These mandatory concepts belong to **SET M**. SET M is a subset of SET V. Semantic Relationships among the concepts of PO framework are discussed in Section 4. These Semantic Relationships belong to set of valid relationships, **SET R**. To run structure and semantic checks using this method is followed:

1. For a concept entered in IPO by research assistants to be valid **(c)** it should be within the scope of SET V and must belong to SET M.
2. For a relationship entered in IPO by research assistants to be valid **(r)** it must belong to SET R.
3. Every tuple (c, r) in IPO belongs to a frameset F. These concepts and relationships are necessary and must be integrated with Original PO.
4. Every tuple ($c^{/}$, r) in IPO belongs to frameset $F^{/}$. Here $c^{/}$ is a concept that does not belong to SET M. These concepts are checked further to see if they belong to SET V. If they do belong to SET V, then the tuple ($c^{/}$, r) is valid and must be integrated with Original PO.
5. All the tuples that do not belong to F and $F^{/}$ are discarded.

Thus, Method 2 is much quicker and efficient way to engineer a trustworthy PO, but it adds to the complexity of the algorithm. The approach proposed here for generating Trustworthy Protein Ontology is currently being implemented to provide a non-redundant, accurate and precise PO framework for future.

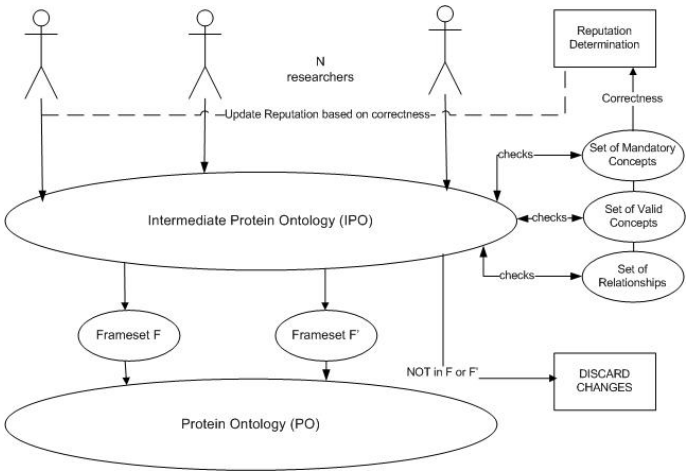

Figure 2: Trustworthy Protein Ontology

To identify the pattern of how correct the information is added we calculate correctness value for every concept entered by the researcher into Intermediate Protein Ontology (IPO). For a correct concept entered the Reputation (**rep**) is increased by 0.1 whereas for a incorrect concept entered the reputation (**rep**) is decreased by 0.05. The final value of reputation for each set of entries by a researcher determines the correctness of information entered by the researchers.

7. Conclusions

The major aim of developing trustworthy ontology is to automate the process to addition and modifications to the Protein Ontology through online interfaces. This also assists in providing a degree of separation between the entered concepts and the actual protein ontology available to the users through the use of Intermediate Protein Ontology (IPO). Trustworthy Protein Ontology Framework makes sure that only valid and correct concepts are added to Protein Ontology. Trustworthy Protein Ontology is under development and systems for addition and editing will be made available through Protein Ontology Website (http://www.proteinontology.info/). In future a an automated reputation based system will provide a means for making changes to be reflected in Protein Ontology based on the reputation of users involved. At the moment, reputation system just provides values to the administrator to modify accesses for the users.

References

[1] R. Corazzon, "Ontology: A Resource Guide for Philosophers," in *http://www.formalontology.it/.* Formalontology, 2003.

[2] A. J. Pretorius, "Lexon Visualisation: Visualising Binary Fact Types in Ontology Bases," in *Semantics Technology and Applications Research Laboratory*, vol. MSc. Brussel: Vrije Universiteit Brussel, 2004.

[3] A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Engineering*. London: Springer, 2003.

[4] S. Schulze-Kremer, "Ontologies for Molecular Biology," presented at Pacific Symposium of Biocomputing, Hawaii, 1998, pp. 693-704.

[5] A. S. Sidhu, T. S. Dillon, and E. Chang, "Current Status of Biomedical Ontologies: Developments in 2006," presented at 2007 IEEE International Conference on Digital Ecosystems and Technologies, Cairns, Australia, 2007, pp. 581-585.

[6] 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.

[7] 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, pp. 465-469.

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

[9] A. S. Sidhu, T. S. Dillon, and E. Chang, "Unification of Protein Data and Knowledge Sources," presented at 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2006), Bournemouth, UK, 2006, pp. 728-737.

[10] 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, pp. 1835-1843.