Protein Ontology Project: 2006 updates

A. S. Sidhu[1], T. S. Dillon[1], B. S. Sidhu[2] & E. Chang[3]
[1]*Faculty of Information Technology, University of Technology Sydney, Australia*
[2]*Punjab State Education Department, India*
[3]*School of Information Systems, Curtin University of Technology, Perth, Australia*

Abstract

Protein Ontology (PO) is a means of formalizing protein data and knowledge; protein ontology includes concepts or terms relevant to the domain, definitions of concepts, and defined relationships between the concepts. PO integrates protein data formats and provides a structured and unified vocabulary to represent protein synthesis concepts. PO provides integration of heterogeneous protein and biological data sources. This paper discusses the updates that happened to Protein Ontology Project since it was last presented at Data Mining 2005 Conference.
Keywords: Protein Ontology, proteomics, bioinformatics, protein informatics, computational proteomics, protein structure, biomedical ontologies, data integration, data semantics.

1 Introduction

Traditional approaches to integrate protein data generally involved keyword searches, which immediately excludes unannotated or poorly annotated data. It also excludes proteins annotated with synonyms unknown to the user. Of the protein data that is retrieved in this manner, some biological resources do not record information about the data source, so there is no evidence of the annotation. An alternative protein annotation approach is to rely on sequence identity, or structural similarity, or functional identification. The success of this method is dependent on the family the protein belongs to. Some proteins have

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060301

high degree of sequence identity, or structural similarity, or similarity in functions that are unique to members of that family alone. Consequently, this approach can't be generalized to integrate the protein data. Clearly, these traditional approaches have limitations in capturing and integrating data for Protein Annotation. For these reasons, we have adopted an alternative method that does not rely on keywords or similarity metrics, but instead uses ontology.

Protein Ontology (PO) is a means of formalizing protein data and knowledge; protein ontology includes concepts or terms relevant to the domain, definitions of concepts, and defined relationships between the concepts. PO integrates protein data formats and provides a structured and unified vocabulary to represent protein synthesis concepts. PO provides integration of heterogeneous protein and biological data sources. PO 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. PO also helps to codify proteomics data for analysis by researchers.

2 PO framework

For developing Protein Ontology (PO) [1–7], we will mainly deal with two main sources of protein annotations: (1) those taken from various protein data sources submitted by authors of protein data themselves from their published experimental results and (2) those that we name annotation that are obtained by an annotator or group of annotators by analysis of raw data (typically a protein sequence or atomic structure description) with various tools extracting biological information from other protein data collections. 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, our protein ontology addresses the following problems of existing protein databases: redundancy, data quality (errors, incorrect annotations, and inconsistencies), lack of standardization in nomenclature etc. Secondly, 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.

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 protein complexes, but does not contain individual protein themselves. The PO database acts as instance store for the PO. 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. 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.

3 PO semantic relationships

Protein Ontology Framework provides specific set of rules to cover these application specific semantics. The rules use only the relationships whose semantics are predefined to establish correspondence among terms in PO. The set of relationships with predefined semantics is: *{SubClassOf, PartOf, AttributeOf, InstanceOf, and ValueOf}*. The PO conceptual modeling encourages the use of strictly typed relations with precisely defined semantics. Some of these relationships *(like SubClassOf, InstanceOf)* are somewhat similar to those in RDF Schema but the set of relationships that have defined semantics in our conceptual PO model is small so as to maintain simplicity of the system. The following is a description of the set of pre-defined semantic relationships in our common PO conceptual model.

SubClassOf: The relationship is used to indicate that one concept is a subclass of another concept, for instance: *SourceCell SubClassOf FunctionalDomains*. That is any instance of SourceCell class is also instance of FunctionalDomains class. All attributes of FunctionalDomains class (_FuncDomain_Family, _FuncDomain_SuperFamily) are also the attributes of SourceCell class. The relationship SubClassOf is transitive.

AttributeOf: This relationship indicates that a concept is an attribute of another concept, for instance: *_FuncDomain_Family AttributeOf Family*. This relationship also referred as PropertyOf, has same semantics as in object-relational databases.

PartOf: This relationship indicates that a concept is a part of another concept, for instance: *Chain PartOf ATOMSequence* indicates that Chain describing various residue sequences in a protein is a part of definition of ATOMSequence for that protein.

InstanceOf: This relationship indicates that an object is an instance of the class, for instance: ATOMSequenceInstance_10 InstanceOf ATOMSequence indicates that ATOMSequenceInstance_10 is an instance of class ATOMSequence.

ValueOf: This relationship is used to indicate the value of an attribute of an object, for instance: *"Homo Sapiens" ValueOf OrganismScientific.* The second concept, in turn has an edge, OrganismScientific AttributeOf Molecule, from the object it describes

In this section we will describe how we used a special semantic relationship like Sequence(s) in Protein Ontology to describe complex concepts defining Structure, Structural Folds and Domains and Chemical Bonds describing Protein Complexes. PO defines these complex concepts as Sequences of simpler generic concepts defined in PO. These simple concepts are Sequences of object and data type properties defining them. A typical example of Sequence is as follows. PO defines a complex concept of ATOMSequence describing three dimensional structure of protein complex as a combination of simple concepts of Chains, Residues, and Atoms as: *ATOMSequence Sequence (Chains Sequence (Residues Sequence (Atoms)))*. Simple concepts defining ATOMSequence are defined as: *Chains Sequence (ChainID, ChainName, ChainProperty); Residues Sequence (ResidueID, ResidueName, ResidueProperty);* and *Atoms Sequence (AtomID, Atom, ATOMResSeqNum, X, Y, Z, Occupancy, TempratureFactor, Element).*

4 Mining PO data

Protein Ontology Database is created as an instance store for various protein data using the PO format. PO provides technical and scientific infrastructure to allow evidence based description and analysis of relationships between proteins. PO uses data sources like PDB, SCOP, OMIM and various published scientific literature to gather protein data. PO Database is represented using OWL. PO Database at the moment contains data instances of following protein families: (1) Prion Proteins, (2) B.Subtilis, (3) CLIC and (4) PTEN. More protein data instances will be added as PO is more developed. The PO instance store at moment covers various species of proteins from bacterial and plant proteins to human proteins. Such a generic representation using PO shows the strength of PO format representation.

We used some standard hierarchical and tree mining algorithms [8] on the PO Database. We compared MB3-Miner (MB3), X3-Miner (X3), VTreeMiner (VTM) and PatternMatcher (PM) for mining embedded subtrees and IMB3-Miner (IMB3), FREQT (FT) for mining induced subtrees of PO Data. In these experiments we are mining Prion Proteins dataset described using Protein Ontology Framework, represented in OWL. For this dataset we map the OWL tags to integer indexes. The maximum height is 1. In this case all candidate subtrees generated by all algorithms would be induced subtrees. Figure 1 shows the time performance of different algorithms.

5 Conclusion

Protein Ontology (PO) discussed in this paper provides a common structured vocabulary for this structured and unstructured information and provides researchers a medium to share knowledge in proteomics domain. It consists of concepts, which are data descriptors for proteomics data and the relations among these concepts. Protein Ontology provides description for protein domains that can be used to describe proteins in any organism.

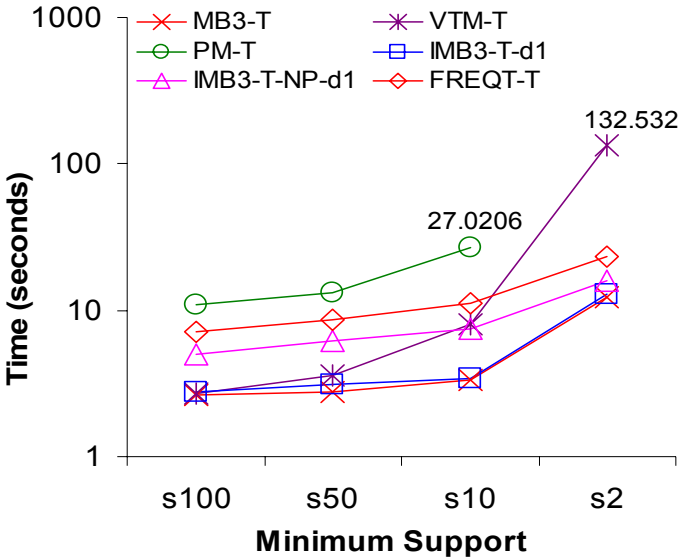

Figure 1: Time performance for Prion dataset of PO data.

References

[1] Sidhu, A. S., T. S. Dillon, et al. (2007). "Knowledge Discovery in Biomedical Data facilitated by Domain Ontologies" in Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data. X. Zhu and I. Davidson. Idea Group Inc.: In Press.

[2] Sidhu, A. S., T. S. Dillon, et al. (2006). Ontology for Data Integration in Protein Informatics. In: Database Modeling in Biology: Practices and Challenges. Z. Ma and J. Y. Chen. New York, NY, Springer Science, Inc.: In Press.

[3] Sidhu, A. S., T. S. Dillon, et al. (2006). Unification of Protein Data and Knowledge Sources. 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2006), Bournemouth, UK, Springer-Verlag. Lecture Notes in Artificial Intelligence (LNAI).

[4] Sidhu, A. S., T. S. Dillon, et al. (2006). Advances in Protein Ontology Project. Special Track on Ontologies for Biomedical Systems at 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, IEEE CS Press.

[5] Hussain, F. K., A. S. Sidhu, et al. (2006). Engineering Trustworthy Ontologies: Case Study of Protein Ontology. Special Track on Ontologies for Biomedical Systems at 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, IEEE CS Press.

[6] Sidhu, A. S., T. S. Dillon, et al. (2005). Ontological Foundation for Protein Data Models. First 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, Springer-Verlag. Lecture Notes in Computer Science (LNCS).

[7] Sidhu, A. S., T. S. Dillon, et al. (2005). Protein Ontology: Vocabulary for Protein Data. 3rd IEEE International Conference on Information Technology and Applications (IEEE ICITA 2005). Sydney, IEEE CS Press. Volume 1: 465-469.

[8] Tan, H., T.S. Dillon, et al. (2006). IMB3-Miner: Mining Induced/Embedded Subtrees by Constraining the Level of Embedding. Accepted for Proceedings of PAKDD 2006.

[9] http://www.proteinontology.info/.