Accomplishments and Challenges of Protein Ontology

Amandeep S. Sidhu[1], Member, IEEE, Tharam S. Dillon[1], Fellow, IEEE, Farookh K. Hussain[2], Elizabeth Chang[2], Member, IEEE

[1]*Faculty of Information Technology, University of Technology Sydney, Australia*
(asidhu, tharam)@ it.uts.edu.au
[2]*School of Information Systems, Curtin University of Technology Perth, Australia*
(Farookh.Hussain, Elizabeth.Chang)@cbs.curtin.edu.au

Abstract

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). 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. We report here our development of PO; a semantic heterogeneity framework based on relationships between PO concepts; and analysis of resultant PO Data of Human Proteins. We also talk in this paper briefly about our ongoing work of designing a trustworthy framework around PO.

1. Introduction

A large number of diverse bioinformatics sources are available today. The future of biological sciences promises more data. No individual data source will provide us with answers to queries that we need to ask. Instead knowledge has to be composed from multiple data sources to answer the queries. Even though multiple databases may cover same data their focus might be different. For example even though Swiss-Prot [1] and PDB [2, 3, 4, and 5] are both protein databases, we might want to get information about sequence as well as structure of a particular protein. In order to answer the query we need to get data about protein from both the sources and combine them in consistent fashion. Bioinformatics researchers have long identified the need of interoperation among protein databases, knowledge bases and other information sources. Despite advances, interoperation among knowledge and data sources is still enabled by hypertext links. Therefore, we need efficient interoperation framework among protein data and information sources.

2. Need for Standards

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 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. Briefly,

Proceedings of the 30th Annual International Computer Software and Applications Conference (COMPSAC'06)
0-7695-2655-1/06 $20.00 © 2006 IEEE

Ontology is a means of formalizing knowledge; at the minimum ontology must include concepts or terms relevant to the domain, definitions of concepts, and defined relationships between the concepts.

In the recent years, several biological data sources have been developed in the biological sciences [6, 7]. These data sources are based on some existing, known conceptual models. Native drivers and wrappers provide access to these data sources and help us restructure the information if needed. In the context of protein data, annotation generally refers to all information about protein other than just protein sequence. In Protein Ontology [8, 9, 10, and 11], we establish application-specific rules - rules that establish correspondence between concepts in different data sources using structured vocabulary of an ontology semi-automatically. The contributions of this work are: (1) Articulation is provided using a pre-defined set of Semantic Relationships, and (2) Query optimization is enabled based on semantic relationships. Our approach has several advantages. Firstly, rule-based articulation generation takes away a lot of effort needed define simple rules during data integration. Secondly, establishment of an interoperation framework enables us to get a better insight on how information is integrated and queries are composed systematically.

3. Protein Ontology Development

The ultimate goal of protein annotator framework or 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.

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. 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. More details about PO Development can be found in [8]. PO Database is represented using XML. At the moment PO Database is constructed semi-automatically, as some of the PO data is entered manually. We are working towards automating the process completely. 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.

4. PO Semantic Framework

4.1 Semantic Relationships

Semantics in protein data is normally not interpreted by annotating systems, since they are not aware of the specific structural, chemical and cellular interactions of protein complexes. 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 SouceCell 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.

AttrributeOf: 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.

4.2 Sequences

Apart from semantic relationships defined in Section 4.1, PO also model relationships like Sequences. By itself semantic relationships described in Section 4.1, does not impose order among the children of the node. In applications using Protein Sequences, the ability of expressing the order is paramount. Generally Protein Sequences are a collection of chains of sequence of residues, and that is the format Protein Sequences have been represented unit now using various data representations and data mining techniques for bioinformatics. When we are defining sequences for semantic heterogeneity of protein data sources using PO we are not only considering traditional representation of protein sequences but also link Protein Sequences to Protein Structure, by linking chains of residue sequences to atoms defining three-dimensional structure. 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)*.

5. PO Data Analysis

We used some standard hierarchical and tree mining algorithms [12] 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 XML. For this dataset we map the XML tags to integer indexes. The maximum height is 1. In this case all candidate subtrees generated by all algorithms would be induced subtrees. Quite interestingly, with Prion dataset of PO the number of frequent candidate subtrees generated is identical for all the major data mining algorithms (Figure 1). This means that the subtrees generated of the PO dataset represented in XML are same for every algorithm.

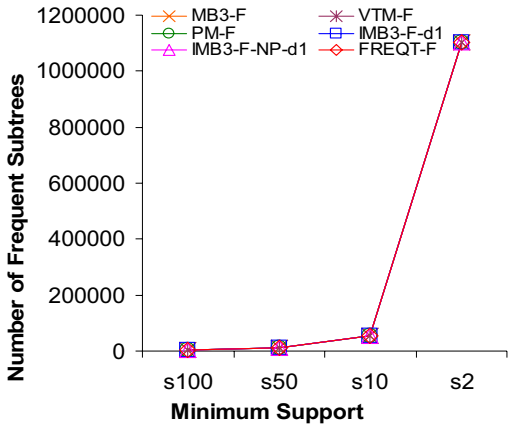

Figure 1: Number of Frequent Subtrees for Prion dataset of PO Data

Therefore the conceptual framework of PO provides a powerful hierarchical classification of concepts, which provides consistency and accuracy in observations of various analysis and reasoning methodologies. This is because the relationships that exist between concepts defined in PO are captured in semantic relationships that assist in composing queries for dynamic data retrieval from distributed protein data sources. We provide specific set of relationships for PO framework to cover data semantics for integrating data information from diverse sources.

6. Comparison

Machine generated protein ontology generated by PRONTO [13] is just a set of terms and relationships between those terms. PRONTO generated ontology does not cover and map all the stages of proteomics process from Protein's Primary Structure to Protein's Quaternary Structure. PRONTO uses iProLink literature mining ontology to search and identify protein names in MEDLINE database of biological literature. It then cross references EBI's UNIPROT database to define relationships between these terms. On the other hand Protein Ontology (PO) integrates data representation frameworks of various protein data sources: PDB, SCOP, RESID and OMIM to provide a unified vocabulary covering all the stages of proteomics process. PRONTO represents only two relationships between the terms of the ontology: is-a relation and part-of relation. Whereas PO represents five different semantic relationships between the terms used in the ontology definition. They are: SubClassOf, PartOf, AttributeOf, InstanceOf, and ValueOf. At the moment we are in the process of developing semantic query algebra over PO conceptual framework to retrieve the data from source databases and populate the XML Database of PO automatically.

PDBML [14] is a XML Schema mapping the PDB Exchange Dictionary. In 2004, we did similar work [15, 16, and 17] to PDBML of creating a XML Schema and RDF Schema mapping of PDB, SWISS-PROT and PIR databases. PDBML lacks the hierarchical relationships as it is licked to logical representation of PDB. The semantics of data is preserved and translation from PDB to XML Schema is simple, but it can't be used to process the content. PO with the power of OWL has no limitations in processing the content.

7. Strengths of PO

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.

PO is represented in Web Ontology Language (OWL). The OWL representation used in Protein Ontology is an XML-Abbrev based (Abbreviated XML Notation), so it can be easily transformed to the corresponding RDF and XML formats without much effort using the available converters. 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.

The PO conceptual modelling encourages the use of strictly typed relations with precisely defined semantics. These relationships will help in defining

semantic query algebra for PO to efficiently reason and query the underlying instance store.

7. Engineering Trustworthy PO

Here we describe 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' Research Assistants. Each of these Research Assistants 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.

7. 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. The limitations of PO in terms of defining new concepts for protein functions and constraints on protein structure does not limit the use of generalized concepts in of PO to define any kind of complex concept for proteomics research in future.

8. References

[1] Boeckmann, B., A. Bairoch, et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Research 31(Database Issue): 365-370.

[2] Berman, H., T. N. Bhat, et al. (2000). "The Protein Data Bank and the challenge of structural genomics." Nature Structural Biology Structural Genomics Supplement(November 2000): 957-959.

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

[4] Weissig, H. and P. E. Bourne (2002). "Protein structure resources." Biological Crystallography D58: 908-915.

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

[6] The Gene Ontology Consortium. Gene ontology: tool for the unification of the biology. Nature Genetics, 25: 25-29, 2000

[7] R. Stevens. Bio-ontology reference collection. http://img.cs.man.ac.uk/stevens/ontopublications.html, 2001.

[8] 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.

[9] Sidhu, A. S., T. S. Dillon, et al. (2006). Protein Ontology Project: 2006 Updates (Invited Paper). Data Mining and Information Engineering 2006. A. Zanasi, C. A. Brebbia and N. F. F. Ebecken. Prague, Czech Republic, WIT Press.

[10] 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).

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] Sidhu, A. S., T. S. Dillon, et al. (2004). Comprehensive Protein Database Representation. 8th International Conference on Research in Computational Molecular Biology 2004 (RECOMB 2004). A. Gramada and P. E. Bourne. San Diego, California., ACM Press: 427-429.