Protein Ontology Project: 2007 updates

A. S. Sidhu, T. S. Dillon & E. Chang
Digital Ecosystems and Business Intelligence Institute,
Curtin University of Technology, Perth, Australia

Abstract

Protein Ontology (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, interactions between two protein molecules and interactions between protein and other macromolecules at cellular level. This paper discusses the updates that happened to the Protein Ontology Project since it was last presented at the Data Mining 2006 Conference.

Keywords: Protein Ontology, proteomics, bioinformatics, protein informatics, computational proteomics, protein structure, biomedical ontologies, data integration, data semantics.

1 Introduction

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 community agree upon concepts underlying the data. 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 [1–4].

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070161

In this paper we discuss the updates that happened to Protein Ontology Project since it was last presented at Data Mining 2006 Conference [5]. Section 2 outlines PO Algebra based on our earlier work on Semantic Relationships in PO [5]. Section 3 discusses results of various data mining algorithms on PO Instance Store. Lastly, Section 4 discusses our future work on building Trustworthy Protein Ontology.

2 PO algebra

Several approaches for data interoperation identified by Karp [6] have been implemented for biological databases. We extend Karp's approach for interoperation not only to protein databases but also to knowledge bases and other information sources. This section outlines algebra for protein data source composition based on our earlier work of Semantic Relationships in Protein Ontology discussed in Data Mining 2006 [5].

The key to scalability of PO conceptual model is the systematic and effective composition of data and information. In this section, we present PO ontology algebra that allows composition of multiple levels of information stored in the ontology for information retrieval. By retaining a log of composition process, we can also, with minimal adaptations, replay the composition whenever any of the underlying data sources that PO integrates change. The algebra has one unary operator: *Select, and three binary operations: Intersection, Union and Difference*.

Select Operator allows us to highlight and select portions of the PO that are relevant to query at hand. Given the PO structure and a concept to be selected, the select operator selects the sub tree rooted at that concept. Given the PO structure and a set of concepts, the select operator selects only those edges in the PO that connect nodes in a given set. Select Operator is defined as:

$OS = \sigma (NS, ES, RS)$ *where*
$NS = Nodes\ (condition = true)$
$ES = Edges\ (\forall N \in NS)$

Intersection Operator is the most important and interesting binary operation. Let *O1 = (N1, E1, R1), and O2 = (N2, E2, R2)* be the two parts of PO whose composition will provide answer to the query submitted by the user. Here N is the set of nodes or concepts of PO, E is the set of edges or the PO hierarchy, and R is set of Semantic Relationships. The intersection of two parts of PO with respect to semantic relationships (SR) of PO is:

$OI\ (1, 2) = O1 \cap_{SR} O2 = (NI, EI, RI)$, *where*
$NI = Nodes\ (SR\ (O1, O2))$,
$EI = Edges\ (E1, NI \cap N1) + Edges\ (E2, NI \cap N2) + Edges\ (SR\ (O1, O2))$, *and*
$RI = Relationships\ (O1, NI \cap N1) + Relationships\ (O2, NI \cap N2) + SR\ (O1, O2) - Edges\ (SR\ (O1, O2))$.

Note that SR is different from R, as it does not include sequences. The nodes in the intersection ontology are those nodes that appear in the semantic relationships, SR. The edges in the intersection ontology are the edges among nodes that are either present in the source parts of the ontology or have been established as a semantic relationship, SR. Relationships in the intersection ontology are the relationships that have not been already been modelled as edges and those relationships present in source parts of the ontology that use only concepts that occur in intersection ontology.

Union Operator combines two parts of the ontology retaining only one copy of the concepts in the intersection. The union of two parts of PO, *O1 = (N1, E1, R1), and O2 = (N2, E2, R2)* with respect to semantic relationships (SR) of PO is expressed as:

$OI\ (1, 2) = O1 \cup_{SR} O2 = (NU, EU, RU)$, *where,*
$NU = N1 \cup N2 \cup NI\ (1, 2)$,
$EU = E1 \cup E2 \cup EI\ (1, 2)$, *and*
$RU = R1 \cup R2 \cup RI\ (1, 2)$, *where,*
$OI\ (1, 2) = O1 \cap_{SR} O2 = (NI\ (1, 2), EI\ (1, 2), RI\ (1, 2))$ *is the intersection of two ontologies.*

The difference of two parts of PO - O1 and O2, computed by **Difference Operator**, written as O1 – O2, includes portions of the first part that are not common to the second part. The difference can be rewritten as $O1 - (O1 \cap_{SR} O2)$. The nodes, edges and relationships that are not in intersection but are present in the first part comprise the difference.

One of the objectives of computing the difference is to optimise the maintenance of PO. As the PO instance store is huge and so many people add instances to it, difference will suggest that instances are not entered properly or there is change in underlying data sources that PO integrates. Change suggested by difference is forwarded to the administrator. If the change happens to be in difference between structures of parts considered, then it does not occur in intersection and is not related to any semantic relationships that establish bridged between the parts of the ontology. Therefore Semantic Relationships do not need to be changed. If the changes is because of changes happened to underlying data sources that PO integrates, then set of concepts and semantic relationships need to be checked for any changes required to remove the difference.

In this section we covered the PO ontology algebra that allows composition of multiple levels of information stored in the protein ontology for information retrieval. The PO approach supports precise composition of information from multiple diverse sources providing semantic relationships between among such sources. This approach allows reliable exploitation of protein information sources without any imposition on the sources themselves. The PO algebra based on semantic relationships allows systematic composition, which unlike integration is more scalable.

3 Mining PO instance store

Tree Mining has attracted lots of interest among the data mining community, due to the increasing use of semi-structured data sources for more meaningful knowledge representations. Here we apply the MB3^{-R} algorithm to the Prions database of PO [7] in order to extract the frequently occurring subtrees. Prions dataset describes Protein Ontology (PO) database for Human Prion proteins in XML format [4, 8]. 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 Figure 1, for varying support thresholds.

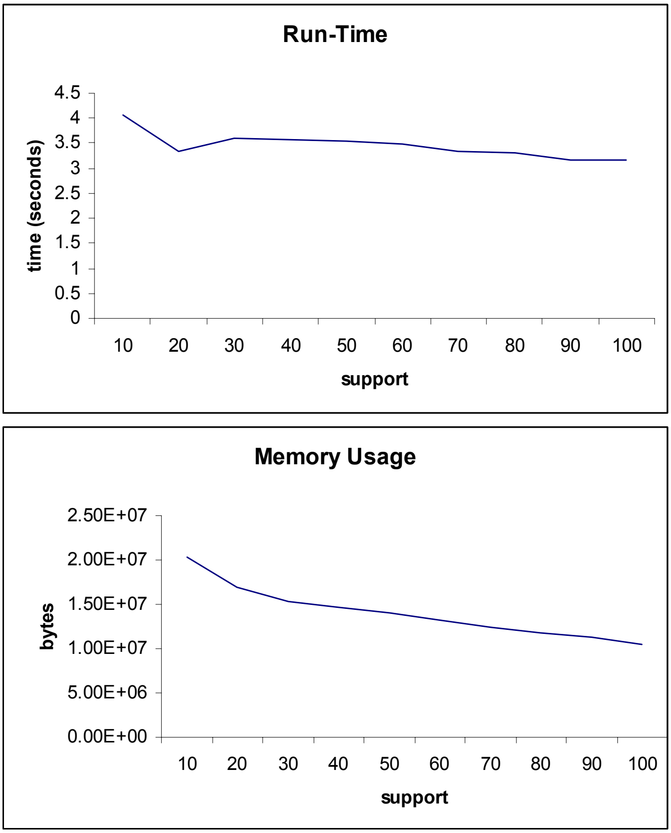

Figure 1: MB3^{-R} run-time and memory usage profile.

We also used some standard hierarchical and tree mining algorithms [9] 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 2 shows the time performance of different algorithms. Our original MB3 has the best time performance for this data.

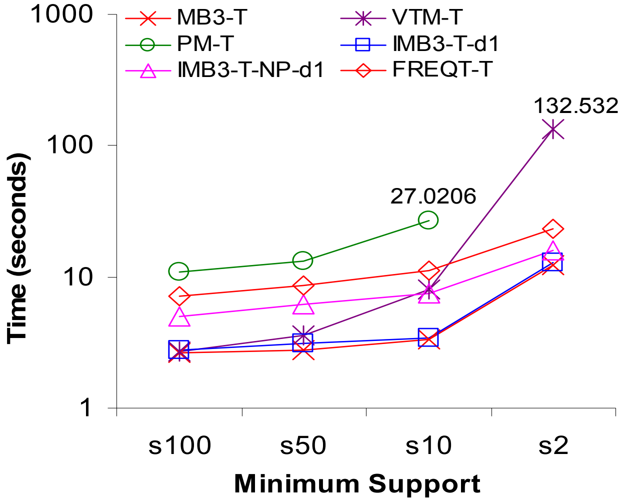

Figure 2: Time Performance for Prion dataset of PO data.

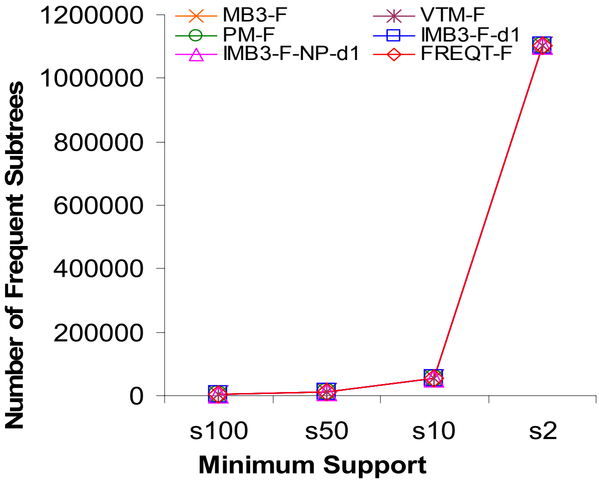

Figure 3: Number of frequent subtrees for Prion dataset of PO data.

Quite interestingly, with Prion dataset of PO the number of frequent candidate subtrees generated is identical for all algorithms (Figure 3). Another observation is that when support is less than 10, PM aborts and VTM performs poorly. The rationale for this could be because the utilized join approach enumerates additional invalid subtrees. Note that original MB3 is faster than IMB3 due to additional checks performed to restrict the level of embedding.

4 Trustworthy Protein Ontology

Here we describe a conceptual framework that we are working on, to engineer Trustworthy Protein Ontology [10]. 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 3. 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.

5 Concluding remarks

Protein Ontology is a part of Standardized Biomedical Ontologies available through the National Center for Biomedical Ontologies [11] along with Gene Ontology [12], Flybase [13], and others (http://cbioapprd.stanford.edu/ncbo/faces/pages/ontology_list.xhtml). More information about Protein Ontology can be found on Protein Ontology Website (http://www.proteinontology.info/). We are in process of adding Protein Ontology to Open Biomedical Ontologies or OBO (http://obo.sourceforge.net/).

Also different research groups are using Protein Ontology for different purposes. Wang et al. [14] shows Protein Ontology as an example of a structured approach for knowledge modeling providing solid inference and retrieval functionalities. Porto [15] discusses Protein Ontology in his report under Ontologies for Bioinformatics. Tan et al. [9] use Protein Ontology generated data set to evaluate their algorithms. Kupfer et al. [16] use Protein Ontology along with Gene Ontology to understand concepts when discussing a coevolution approach for database schemas. Bolshakova et al. [17] discuss protein ontology under a section on Biomedical Ontologies while comparing data based and ontology based approaches for cluster validation of microarrays.

IQlue [18] references Protein Ontology when reviewing Ontology Development in their white paper. Dhanapalan and Chen [19] discuss protein ontology in detail when doing case study of integrating protein interaction data using semantic web technology. Pinagé and Brilhante [20] used Protein Ontology for Protein Structure Homology Modeling. Just Recently researchers [21] 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. Kupfer et al. [22] reuses the concept of chains from Protein Ontology when proposing database ontology for signal transduction pathways. Lastly, Lacroix et al. [23] discuss Protein Ontology briefly when proposing a semantic model to integrate biological resources.

References

[1] 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, pp. 588-592.

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

[3] 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.

[4] 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, pp. 51-60.

[5] A. S. Sidhu, T. S. Dillon, B. S. Sidhu, and E. Chang, "Protein Ontology Project: 2006 Updates," presented at 7th Data Mining and Information Engineering 2006 (Data Mining 2006), Prague, Czech Republic, 2006, pp. 301-306.

[6] P. Karp, "Database links are a foundation for interoperability," *Trends in Biotechnology*, vol. 14, pp. 273-279, 1996.

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

[8] 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.

[9] H. Tan, T. S. Dillon, F. Hadzic, E. Chang, and L. Feng, "IMB3-Miner: Mining Induced/Embedded Subtrees by Constraining the Level of

Embedding," presented at 10th Pacific-Asia Knowledge Discovery and Data Mining Conference (PAKDD 2006), Singapore, 2006, pp. 450-461.

[10] F. K. Hussain, A. S. Sidhu, T. S. Dillon, and E. Chang, "Engineering Trustworthy Ontologies: Case Study of Protein Ontology," presented at 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, 2006, pp. 617-622.

[11] 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.

[12] 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.

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

[14] 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, pp. 13-18.

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

[16] 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, pp. 605-610.

[17] 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, pp. 539-543.

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

[19] 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.

[20] 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

[21] 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

[22] 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.

[23] 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, pp. 63-73.