Integration of Protein Data Sources through PO

Amandeep S. Sidhu[1], Tharam S. Dillon[1], and Elizabeth Chang[2]

[1]Faculty of Information Technology, University of Technology, Sydney, Australia
{asidhu, tharam}@it.uts.edu.au

[2]School of Information Systems, Curtin University of Technical University, Perth, Australia
Elizabeth.Chang@cbs.curtin.edu.au

Abstract. Resolving heterogeneity among various protein data sources is a crucial problem if we want to gain more information about proteomics process. Information from multiple protein databases like PDB, SCOP, and UniProt need to integrated to answer user queries. Issues of Semantic Heterogeneity haven't been addressed so far in Protein Informatics. This paper outlines protein data source composition approach based on our existing work of Protein Ontology (PO). The proposed approach enables semi-automatic interoperation among heterogeneous protein data sources. The establishment of semantic interoperation over conceptual framework of PO enables us to get a better insight on how information can be integrated systematically and how queries can be composed. The semantic interoperation between protein data sources is based on semantic relationships between concepts of PO. No other such generalized semantic protein data interoperation framework has been considered so far.

1. Introduction

In accelerating quest for disease biomarkers, the use of high-throughput technologies, such as DNA microarrays and proteomics experiments, has produced vast datasets identifying thousands of genes whose expression patterns differ in diseased versus normal samples. Although many of these differences may reach statistical significance, they are not biologically meaningful. For example, reports of mRNA or protein changes of as little as two-fold are not uncommon, and although some changes of this magnitude turn out to be important, most are attributes to disease-independent differences between the samples. Evidence gleaned from other studies linking genes to disease is helpful, but with such large datasets, a manual literature review is often not practical. The power of these emerging technologies – the ability to quickly generate large sets of data – has challenged current means of evaluating and validating these data. Thus, one important example of a data rich but knowledge poor area is biological sequence mining. In this area, there exist massive quantities of data generated by the data acquisition technologies. The bioinformatics solutions addressing these data are a major current challenge. However, domain specific ontologies such as Gene Ontology [1], MeSH [2] and Protein Ontology (PO) [3, 4, 5, and 6] exist to provide context to this complex real world data.

2. Protein Ontology Conceptual Framework

Advances in technology and the growth of life sciences are generating ever increasing amounts of data. High-throughput techniques are regularly used to capture thousands of data points in an experiment. The results of these experiments normally end up in scientific databases and publications. Although there have been concerted efforts to capture more scientific data in specialist databases, it is generally acknowledged that only 20 per cent of biological knowledge and data is available in a structured format. The remaining 80 per cent of biological information is hidden in the unstructured scientific results and texts. Protein Ontology (PO) [3, 4, 5, and 6] 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 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. Protein Ontology provides description for protein domains that can be used to describe proteins in any organism. Protein Ontology Framework describes: (1) Protein Sequence and Structure Information, (2) Protein Folding Process, (3) Cellular Functions of Proteins, (4) Molecular Bindings internal and external to Proteins and (5) Constraints affecting the Final Protein Conformation. Protein Ontology uses all relevant protein data sources of information. The structure of PO provides the concepts necessary to describe individual proteins, but does not contain individual protein themselves. A database based on PO acts as instance store for the PO. PO uses data sources include new proteome information resources like PDB, SCOP, and RESID as well as classical 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. PO Database is represented using XML. 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 Complete Class Hierarchy of Protein Ontology (PO) is shown in **Figure 1**. More details about PO is available at the website: **http://www.proteinontology.info/**

- ***ProteinOntology***
  - ***AtomicBind***
  - ***Atoms***
  - ***Bind***
  - ***Chains***
  - ***Family***
  - ***ProteinComplex***
    - ***ChemicalBonds***
      - ***CISPeptide***
      - ***DisulphideBond***
      - ***HydrogenBond***
      - ***ResidueLink***
      - ***SaltBridge***
    - ***Constraints***
      - ***GeneticDefects***
      - ***Hydrophobicity***
      - ***ModifiedResidue***
    - ***Entry***
      - ***Description***
      - ***Molecule***
      - ***Reference***
    - ***FunctionalDomains***
      - ***ActiveBindingSites***
      - ***BiologicalFunction***
        - ***PathologicalFunctions***
        - ***PhysiologicalFunctions***
      - ***SourceCell***
    - ***StructuralDomains***
      - ***Helices***
        - ***Helix***
          - ***HelixStructure***
      - ***OtherFolds***
        - ***Turn***
          - ***TurnStructure***
      - ***Sheets***
        - ***Sheet***
          - ***Strands***
    - ***Structure***
      - ***ATOMSequence***
      - ***UnitCell***
  - ***Residues***
  - ***SiteGroup***

Figure 1. Class Hierarchy of Protein Ontology

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.

3. Comparing GO and PO

Gene Ontology (GO) [1] defines a structured controlled vocabulary in the domain of biological functionality. GO initially consisted of a few thousand terms describing the genetic workings of three organisms and was constructed for the express purpose of database interoperability; it has since grown to a terminology of nearly 16,000 terms and is becoming a de facto standard for describing functional aspects of biological entities in all types of organisms. Furthermore, in addition to (and because of) its wide use as a terminological source for database-entry annotation, GO has been used in a wide variety of biomedical research, including analyses of experimental data [1] and predictions of experimental results [7]. Characteristics of GO that we believe are most responsible for its success: community involvement; clear goals; limited scope; simple, intuitive structure; continuous evolution; active curation; and early use.

It is clear that organisms across the spectrum of life, to varying degrees, possess large numbers of gene products with similar sequences and roles. Knowledge about a given gene product (i.e., a biologically active molecule that is the deciphered end product of the code stored in a gene) can often be determined experimentally or inferred from its similarity to gene products in other organisms. Research into different biological systems uses different organisms that are chosen because they are amenable to advancing these investigations. For example, the rat is a good model for the study of human heart disease, and the fly is a good model to study cellular differentiation. For each of these model systems, there is a database employing curators who collect and store the body of biological knowledge for that organism. This enormous amount of data can potentially add insight to related molecules found in other organisms. A reliable wet-lab biological experiment performed in one organism can be used to deduce attributes of an analogous (or related) gene product in another organism, thereby reducing the need to reproduce experiments in each individual organism (which would be expensive, time-consuming, and, in many organisms, technically impossible). Mining of Scientific Text and Literature is done to generate list of keywords that is used as GO terms. However, querying heterogeneous, independent databases in order to draw these inferences is difficult: The different database projects may use different terms to refer to the same concept and the same terms to refer to different concepts. Furthermore, these terms are typically not formally linked with each other in any way. GO seeks to reveal these underlying biological functionalities by providing a structured controlled vocabulary that can be used to describe gene products, and shared between biological databases. This facilitates querying for gene products that share biologically meaningful attributes, whether from separate databases or within the same database.

Challenges faced while developing GO from unstructured and structured data sources are addressed while developing PO. Protein Ontology is a conceptual model that aim to support consistent and unambiguous knowledge sharing and that provide a framework for protein data and knowledge integration. PO links concepts to their interpretation, i.e. specifications of their meanings including concept definitions and relationships to other concepts. Apart from semantic relationships defined in Section 2, PO also model relationships like Sequences. By itself semantic relationships described in Section 2, 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)*. Semantic Interoperability Framework used in PO is depicted **Figure 2**.

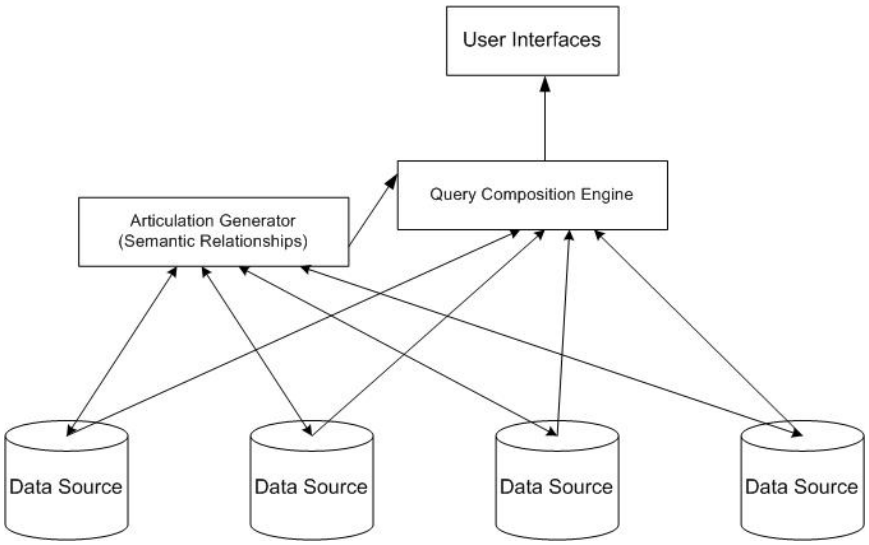

Figure 2. Semantic Interoperability Framework for PO

Therefore, PO reflects the structure and relationships of Protein Data Sources. PO removes the constraints of potential interpretations of terms in various data sources and provides a structured vocabulary that unifies and integrates all data and knowledge sources for proteomics domain **(Figure 3)**. There are seven subclasses of Protein Ontology (PO), called Generic Classes that are used to define complex concepts in other PO Classes: Residues, Chains, Atoms, Family, AtomicBind, Bind, and SiteGroup. Concepts from these generic classes are reused in various other PO Classes for definition of Class Specific Concepts. Details and Properties of Residues in a Protein Sequence are defined by instances of Residues Class. Instances of Chains of Residues are defined in Chains Class. All the Three Dimensional Structure Data of Protein Atoms is represented as instances of Atoms Class. Defining Chains, Residues and Atoms as individual classes has the benefit that any special properties or changes affecting a particular chain, residue and atom can be easily added. Protein Family class represents Protein Super Family and Family Details of Proteins. Data about binding atoms in Chemical Bonds like Hydrogen Bond, Residue Links, and Salt Bridges is entered into ontology as an instance of AtomicBind Class. Similarly the data about binding residues in Chemical Bonds like Disulphide Bonds and CIS

Peptides is entered into ontology as an instance of Bind Class. All data related to site groups of the active binding sites of Proteins is defined as instances of SiteGroup Class. 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.

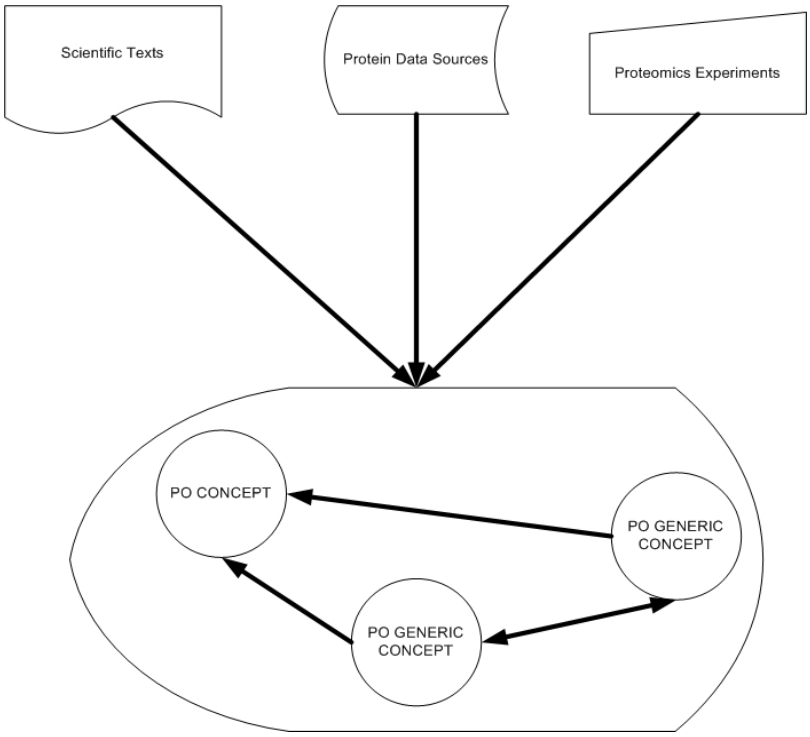

Figure 3. Unification of Protein Data and Knowledge

As such PO can be used to support automatic semantic interpretation of data and knowledge sources, thus providing a basis for sophisticated mining of information.

4. Mining facilitated by Protein Ontology

The 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 XML. 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 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. **Figure 4** shows the time performance of different algorithms. Our original MB3 has the best time performance for this data.

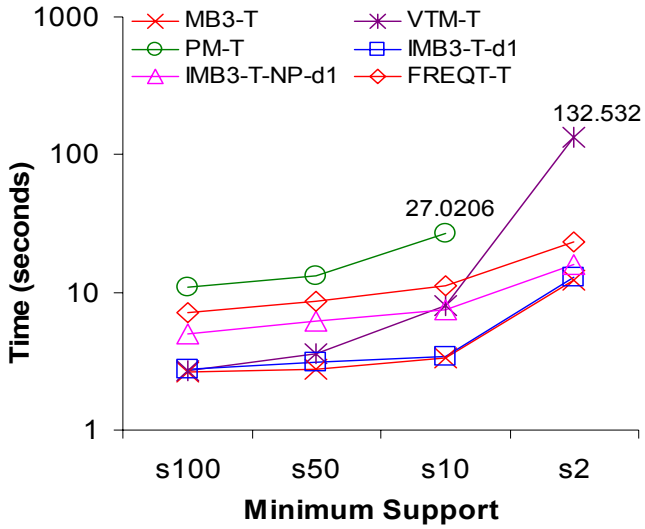

Figure 4. Time Performance for Prion dataset of PO Data

Quite interestingly, with Prion dataset of PO the number of frequent candidate subtrees generated is identical for all algorithms **(Figure 5)**. 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.

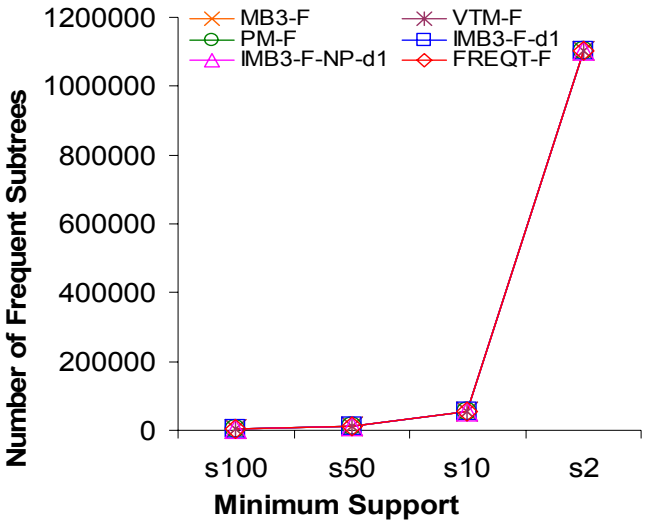

Figure 5. Number of Frequent Subtrees for Prion dataset of PO Data

5. Conclusion

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

References

[1] GO Consortium (2001). "Creating the Gene Ontology Resource: Design and Implementation." Genome Research 11: 1425-1433.

[2] Nelson, Stuart J.; Schopen, Michael; et al. (2004).The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. In: Fieschi, M. et al., editors. Proceedings of the 11th World Congress on Medical Informatics; 2004 Sep 7-11; San Francisco, CA. Amsterdam: IOS Press; pp. 67-69.

[3] 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.

[4] 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.

[5] 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).

[6] 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.

[7] GO Consortium and S. E. Lewis (2004). "Gene Ontology: looking backwards and forwards." Genome Biology 6(1): 103.1-103.4.

[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.