21st IEEE International Symposium on Computer-Based Medical Systems

Use of Protein Ontology to Enable Data Exchange for Complex Proteomic Experiments

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

[1]Digital Ecosystems and Business Intelligence Institute,
Curtin University of Technology Perth, Australia
(Amandeep.Sidhu, Tharam.Dillon, Elizabeth.Chang)@cbs.curtin.edu.au
[2]Department of Computer Science and Computer Engineering,
La Trobe University, Melbourne, Australia

Abstract

Ontologies and controlled vocabularies are being established by many groups to provide roadmaps through the confused mass of data currently being generated from increasingly large-scale experimental biological experiments. The world of protein chemistry is no exception to this rule, with Protein Ontology (PO) having lead the field by providing a framework in which individual molecules and complexes can be defined by their structure, function and cellular location. PO performs searches across the protein databases using a standard nomenclature consistent to all entries.

1. Introduction

Proteomics is often described as the study of the protein translation products of the genome of a given organism but, in reality, this definition should be expanded to an understanding of the expression pattern and state of all proteins transcribed under a given set of conditions and the alteration of these parameters in response to a specific change to these conditions. The proteome of a cell encompasses the identity, subcellular location, post-translational modifications and protein-protein interactions made by the spectra of proteins expressed at any one moment in time and also how all these effect the function of both an individual protein and the cell as a whole. In order to map this, a multitude of experimental techniques have been developed. Proteins have first to be isolated and separated from a given biological sample, the latter usually either by 2-dimensional gel electrophoresis or by HPLC. The analytes are then ionized in the gas phase and the mass of the resulting peptide fragments measured by mass spectrometry. Such analyses will provide an expression map of the protein content of the cell under the defined experimental conditions. Further techniques have been developed to provide further detail of the state of these proteins and their actual location within the cell. To fully understand the biological processes and pathways in which any one protein molecule may be involved, it is necessary to be aware of the interactions that molecule makes with other proteins, nucleic acids and small molecules within the cell.

Multiple laboratories have contributed data to various databases such as the Human Proteome Organization (HUPO) Plasma Proteome Project and Human Plasma Peptide Atlas. Between laboratories, the identification of proteins can be difficult to cross-validate based on the diversity of experimental approaches and measurements, and the different databases do not generally overlap in each of their identified sets of proteins. Multiple protein databases may cover same data but 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. Ontologies in Proteomics

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

1063-7125/08 $25.00 © 2008 IEEE
DOI 10.1109/CBMS.2008.145

conceptual models. Native drivers and wrappers provide access to these data sources and help us restructure the information if needed. Ontologies and controlled vocabularies are being established by many groups to provide roadmaps through the confused mass of data currently being generated from increasingly large-scale experimental biological experiments. The world of protein chemistry is no exception to this rule, with Protein Ontology (PO) having lead the field by providing a framework in which individual molecules and complexes can be defined by their structure, function and cellular location. PO performs searches across the protein databases using a standard nomenclature consistent to all entries.

In Protein Ontology [8, 9, 10, and 11], we define relationships that establish correspondence between concepts in different data sources using structured vocabulary of ontology semi-automatically. 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. Protein Ontology Database is created as an instance store for various protein data using the PO format. PO uses data sources like PDB, SCOP, OMIM and various published scientific literature to gather protein data. Such a generic representation using PO shows the strength of PO format representation.

3. PO Algebra for Data Exchange

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 modelling 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. Case Study: Bacillus Subtilis

Bacillus Subtilis is a bacterium commonly found in soil. B. Subtilis has the ability to form a tough, protective endospore, allowing the organism to tolerate extreme environmental conditions. B. Subtilis has proven to be highly amenable to genetic manipulation, and has therefore become widely adopted as a model organism for laboratory studies, especially of sporulation, which is a simplified example of cellular differentiation. It is also heavily flagellated, which gives B. Subtilis the ability to move quite quickly.

With the assistance of our colleagues we have populated the PO Instance Store with all the major protein complexes that belong to the B. Subtilis family. All the PO Instance files are available for download at the PO website (http://proteinontology.info/proteins.htm).

In this section, we examine how information is integrated from various data sources for a B. Subtilis Protein using Protein Ontology. We will integrate information about a B. Subtilis septum formation MAF

protein complex with D - (UTP) from Protein Data Bank and UniProt in this example.

Title, Keywords, Experimental Method (EXPDTA), and Authors from PDB are described in Description Concept of Protein Ontology. Information about molecules present in the protein complex from PDB is given in the Molecule Concept of Protein Ontology. Organism and Cellular Source where protein resides from PDB is described using Source Cell Concept of Protein Ontology. Literature References for Protein Complex from PDB are described using Reference Concept of Protein Ontology. The Atom Record (Atom ID: 583) is described in Protein Ontology. Although the Protein Data Bank (PDB) is very comprehensive, it does not provide information about Protein Sequences in traditional FASTA format, which is still widely used, and provides no information about the basic functionality of the protein complex. We try to gather this information from UniProt database and integrate it with information gathered from PDB in Protein Ontology. Protein Sequence information from UniProt is integrated with Protein Structure information from PDB using ATOMSequence Concept in Protein Ontology. ATOMSequence is constructed using generic concepts of Chains, Residues, and Atoms. The reasoning is already there in the underlying Protein Data, as each Chain in a Protein represents a sequence of Residues, and each Residue is defined by a number of three-dimensional atoms in the Protein Structure. Information about the functionality of B. Subtilis Protein is described in Protein Ontology as a Physiological Function. More detailed discussion can be found in our related publication [9].

5. Future Work

Although the Protein Ontology (PO) provides a unified vocabulary for capturing declarative knowledge about the protein domain and classifies it in a rich hierarchy of concepts and their relationships, it is still a passive structure. This passive nature arises from the fact that the user has to approach the ontology with enough knowledge to know what to look for. We move beyond this point and in future we propose to create an active support using the protein ontology as the background. In our future work, we will create agents that will interact and mediate between the ontology and human agents. The innovation lies in the fact that we are creating an active support for biologists, so that when a biologist uses the web service, he or she will obtain an advisor or a recommender rather than just an organization of the knowledge.

For example, if a biologist knows the terms, he or she can go ahead and look for information from the body of knowledge in protein ontology concepts to obtain a clarification of the terms. If he or she wants to know the relationships between the terms, he can use the Conceptual Hierarchy of Protein Ontology that relates concepts using semantic relationships. On the other hand, if he or she wants advice on what to do to find a certain piece of information from protein ontology, then we need some active components, which are intelligent enough to utilise the protein ontology framework to advise the user. It is the creation of these active components that is the subject of our future work.

In order to achieve this, we will have to extend our existing knowledge of the interaction between agents and ontologies and between agents and biologists. However, we will do it purely in the context of supporting the Protein Ontology. This approach can be used in a generic sense for any ontology, and the result of this project should be capable of being used in any other biomedical ontologies. We will develop a multi-agent based approach that will integrate the protein ontology and expertise of biologists for bioinformatics communities. The integration of the intelligent agents and the protein ontology is an innovative technology that can significantly improve the recommender approach for the multi-site distributed development of protein data sources through protein ontology.

In particular, protein ontology enables an active ecology of agents to convey, consume and act on protein data and information autonomously, according to domain knowledge provided by it. The approach has the ability to collect data and information from diverse protein data sources and scientific literature for current issues in proteomics. We will develop a multi-agent system based on the protein ontology framework present in this thesis that facilitates meaningful communication, discussion, negotiation and information exchange between protein data and information sources through collaborative agents. These collaborative agents will further enhance the semantic interoperability among protein data sources provided by protein ontology. This intelligent agent architecture will also enable the development of an approach for the sharing of expertise, documents and progress of the protein data sources.

6. Conclusions

The design and use of ontologies to describe experimental data and enable the storage and exchange of proteomics data in a format that allows subsequent users a clear and comprehensive understanding of proteomics data and experiments. Protein Ontology uses a structured nomenclature to describe data formats, molecular interactions, the features on a molecule responsible for such interactions and the experimental methods by which both interaction and features were determined.

7. 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., Dillon, T.S. and Chang, E. (2007) Protein Ontology. In Chen, J. and Sidhu, A.S. (eds), Biological Database Modeling. Artech House, New York.

[9] Sidhu, A.S., Dillon, T.S. and Chang, E. (2007) Protein Ontology Instance Store. 3rd IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2006) in conjunction with OTM 2007. Springer-Verlag, Portugal.

[10] Sidhu, A.S., Dillon, T.S. and Chang, E. (2007) Deployment of Protein Ontology Framework, International Journal of Aplied Bioengineering, 1, 34-41.

[11] Sidhu, A.S., Dillon, T.S., Sidhu, B.S. and Setiawan, H. (2004) Protein Knowledge Meta Model, Molecular & Cellular Proteomics, 10, S262-S263.