Protein Data Integration through Ontologies

Amandeep S. Sidhu
Digital Ecosystems and Business Intelligence Institute,
Curtin University of Technology, Perth, Australia
a.sidhu@curtin.edu.au

Abstract. In this paper, we consider the challenges of information integration in proteomics from the prospective of researchers using information technology as an integral part of their discovery process. Specifically, data integration, meta-data specification, data provenance and data quality are discussed here. Here we review existing protein data integration methods and propose the use of common vocabulary for protein data integration using ontologies.

1. Introduction

The advent of automated and high-throughput technologies in biological research and the progress in the genome projects has led to an ever-increasing rate of data acquisition and exponential growth of data volume. However, the most striking feature of data in life science is not its volume but its diversity and variability. The biological data sets are intrinsically complex and are organised in loose hierarchies that reflect our understanding of complex living systems, ranging from genes and proteins, to protein-protein interactions, biochemical pathways and regulatory networks, to cells and tissues, organisms and populations, and finally ecosystems on earth. This system spans many orders of magnitudes in time and space and poses challenges in informatics, modelling, and simulation that goes beyond any scientific endeavour. Reflecting the complexity of biological systems, the types of biological data are highly diverse. They range from plain text of laboratory records and literature publications, nucleic acid and protein sequences, three-dimensional atomic structure of molecules, and biomedical images with different levels of resolutions, to various experimental outputs from technology as diverse as microarray chips, light and electronic microscopy, Nuclear Magnetic Resonance (NMR), and mass spectrometry. This presents a great challenge in modelling biological objects. In this paper we will discuss existing protein data integration methods and propose the use of common vocabulary for protein data integration using ontologies.

2. Related Works

2.1 Existing Data Integration Methodologies

In the context of protein data, annotation generally refers to all information about protein other than protein sequence. 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 generalised to integrate the protein data. Clearly, these traditional approaches have limitations in capturing and integrating data for Protein Annotation. Perhaps these problems could be addressed more easily in the context of a more general logical structure. For these reasons, we have adopted an alternative method that does not rely on keywords or similarity metrics, but instead uses ontology. Ontology is a means of formalising knowledge; at the minimum ontology must include concepts or terms relevant to the domain, definitions of concepts, and defined relationships between the concepts.

2.2 Need for Biomedical Ontologies

Semantics of protein data is usually hard to define precisely because they are not explicitly stated but are implicitly included in database design. Proteomics is not a single, consistent domain; it is composed of various smaller focused research communities, each having a different data format. Data Semantics would not be a significant issue if researchers only accessed data from within a single research domain, but this is not usually the case. Typically, researchers require integrated access to data from multiple domains, which requires resolving terms that have slightly different meanings across communities.

To integrate the data generated through a web-based system the research results need to be consistent, classified, retrieved and queried using a unified common vocabulary. This will facilitate sharing and cross-linkage of the results and will provide mechanism for interoperability between various databases. We note most of the modern biological databases use data descriptors specified in a schema curated according to the needs and requirements of the immediate community, without consideration to interoperability with other databases. This underlying issue of heterogeneity in biological domain can be addressed partly by developing a common vocabulary using Ontologies for data modelling and knowledge sharing as demonstrated in Genomics by Gene Ontology (Lewis, 2004, Ashburner et al., 2001) and MGED Ontology (Whetzel et al., 2006).

The Gene Ontology is a collaborative effort to create a controlled vocabulary of gene and protein roles in cells, addressing the need for consistent descriptions of gene products in different databases. The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. One of the important uses of GO is the prediction of gene function based on patterns of annotation. For example, if annotations for two attributes tend to occur together in the database, then the gene holding one attribute is likely to hold for the other as well (King et al., 2003). In this way, functional predictions can be made by applying prior knowledge to infer the function of the new entity (either a gene or a protein).

The MGED Ontology (MO) is developed by the Microarray Gene Expression Data (MGED) Society. MO provides terms for annotating all aspects of a microarray experiment from the design of the experiment and array layout, through to preparation of the biological sample and protocols used to hybridise the RNA and analyse the data. MO is a species-neutral ontology that focuses on commonalities among experiments rather than differences between them. MO is primarily an ontology used to annotate microarray experiments; however, it contains concepts that are universal to other types of functional genomics experiments. The major component of the ontology involves biological descriptors relating to samples or their processing.

3. Protein Ontology (PO)

We built the Protein Ontology (Sidhu et al., 2007; Sidhu et al., 2005; Sidhu et al., 2004) to integrate protein data formats and provide a structured and unified vocabulary to represent protein synthesis concepts. Following PO's lead recently two other ontologies have been designed for as well. PRotein Ontology (PRO) (Natale et al., 2007) to facilitate protein annotation and to guide new experiments. The components of PRO extend from the classification of proteins on the basis of evolutionary relationships to the representation of the multiple protein forms of a gene. Proteomics Process Ontology (ProPreO) (Sahoo et al., 2006) enables a detailed description of proteomics experimental processes and data.

Protein Ontology (PO) provides an integration of heterogeneous protein and biological data sources. It converts the enormous amounts of data collected by geneticists and molecular biologists into information that scientists, physicians and other health care professionals. PO consists of concepts, which are data descriptors for proteomics data and the relationships among these concepts. PO has (1) a hierarchical classification of concepts, from general to specific; (2) a list of properties related to each concept; (3) a set of relationships to link concepts in ontology in more complicated ways then implied by the hierarchy; and (4) a set of algebraic operators for querying protein ontology instances. In this section, we will briefly discuss various concepts and relationships that make up PO. More details about Protein Ontology are available on the website (http://proteinontology.org.au/).

3.1 Protein Ontology Concepts

The root concept in PO is *ProteinOntology*. For each instance of protein that is entered into PO, the submission information is entered for *ProteinOntology* concept. There are seven concepts of PO, called **Generic Concepts** that are used to define complex PO Concepts: *{Residues, Chains, Atoms, Family, AtomicBind, Bind, and SiteGroup}*. These generic concepts are reused in defining complex PO concepts. We now briefly describe these generic concepts. Details and Properties of Residues in a Protein Sequence are defined by instances of *Residues* concept. Instances of Chains of Residues are defined in *Chains* concept. All the Three Dimensional Structure Data of Protein Atoms are represented as instances of *Atoms* concept. Defining Chains, Residues and Atoms as individual concepts has the advantage that any special properties or changes affecting a particular chain, residue and atom can be easily

added. *Family* concept represents Protein Super Family and Family Details of Proteins. Data about binding atoms in Chemical Bonds like Hydrogen Bond, Residue Links, and Salt Bridges are entered into ontology as an instance of *AtomicBind* concept. Similarly, the data about binding residues in Chemical Bonds like Disulphide Bonds and CIS Peptides is entered into ontology as an instance of *Bind* concept. When defining the generic concepts of *AtomicBind* and *Bind* in PO we again reuse the generic concepts of *Chain*, *Residue*, and *Atom*. All data related to site groups of the active binding sites of Proteins are defined as instances of *SiteGroup* concept. In PO, the notions classification, reasoning, and consistency are applied by defining new concepts from the defined generic concepts.

The Main Concept for definition of Protein Complexes in the Protein Ontology is *ProteinComplex*. *ProteinComplex* concept defines one or more Proteins in the Complex Molecule. Six sub concepts of ProteinComplex: *Entry, Structure, StructuralDomains, FunctionalDomains, ChemicalBonds, and Constraints* provide a complete understanding of the sequence, structure and functional interactions of proteins. They define sequence, structure, function, and chemical bindings of the Protein Complex and are **derived concepts** formed from the generic concepts discussed earlier.

PO describes Protein Complex Entry and the Molecules contained in Protein Complex are described using *Entry* concept and its sub-concepts of *Description*, *Molecule* and *Reference*. Molecule reuses the generic concept of *Chain* to represent the linkage of molecules in the protein complex to the chain of residue sequences.

Protein Sequence and Structure data are described using *Structure* concept in PO with sub-concepts *ATOMSequence* and *UnitCell*. *ATOMSequence* represents protein sequence and structure and is made of generic concepts of *Chain*, *Residue* and *Atom*. Protein Crystallography Data is described using the *UnitCell* concept.

Protein Structural Folds and Domains are defined in PO using the derived concept of *StructuralDomains*. Family and Super Family of the organism in which protein is present are represented in *StructuralDomains* by reference to the generic concept of *Family*. Structural Folds in protein are represented by sub-concepts of *Helices*, *Sheets* and *Other Folds*. Each definition of structural folds and domains also reuses the generic concepts of *Chain* and *Residue* for describing the Secondary Structure of Proteins.

PO has the first Functional Domain Classification Model for proteins defined using the derived concept of *FunctionalDomains*. Like *StructuralDomains*, the Family and Super Family of the organism in which protein is present, are represented in *FunctionalDomains* by reference to the generic concept of *Family*. *FunctionalDomains* describes the Cellular and Organism Source of Protein using *SourceCell* sub-concept, Biological Functionality of Protein using *BiologicalFunction* sub-concept, and describes Active Binding Sites in Protein using *ActiveBindingSites* sub-concept. Active Binding Sites are represented in PO as a collection of various Site Groups, defined using *SiteGroup* generic concept.

Various chemical bonds used to bind various substructures in a complex protein structure are defined using *ChemicalBonds* concept in PO. Chemical Bonds are defined by their respective sub-concepts are: *DisulphideBond, CISPeptide,*

HydrogenBond, ResidueLink, and SaltBridge. They are defined using generic concepts of *Bind* and *Atomic Bind*.

Various constraints that affect the final protein structural conformation are defined using the *Constraints* concept of PO. The constraints described in PO at the moment are: Monogenetic and Polygenetic defects present in genes that are present in molecules making proteins described using *GeneticDefects* sub-concept, Hydrophobic properties of proteins described using *Hydrophobicity* sub-concept, and Modification in Residue Sequences are described using in *ModifiedResidue* sub-concept.

3.2 Relationships 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. A Protein Ontology Framework provides a specific set of rules to cover these application specific semantics. The rules use only the relationships whose semantics are predefined in PO to establish correspondence among terms. 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 (W3C-RDFSchema 2004) but the set of relationships that have defined semantics in our conceptual PO model is too small to maintain the simplicity of the model. The following is a brief description of the set of pre-defined semantic relationships in our common PO conceptual model. *SubClassOf* relationship is used to indicate that one concept is a specialisation of another concept. *AttributeOf* relationship indicates that a concept is an attribute of another concept. *PartOf* relationship indicates that a concept is a part of another concept. *InstanceOf* relationship indicates that an object is an instance of the concept. *ValueOf* relationship is used to indicate the value of an attribute of an object. By themselves, the relationships described above do not impose order among the children of the node. We defined a special relationship called *Sequence(s)* in PO to describe and impose order in complex concepts defining Structure, Structural Folds and Domains and Chemical Bonds of Proteins.

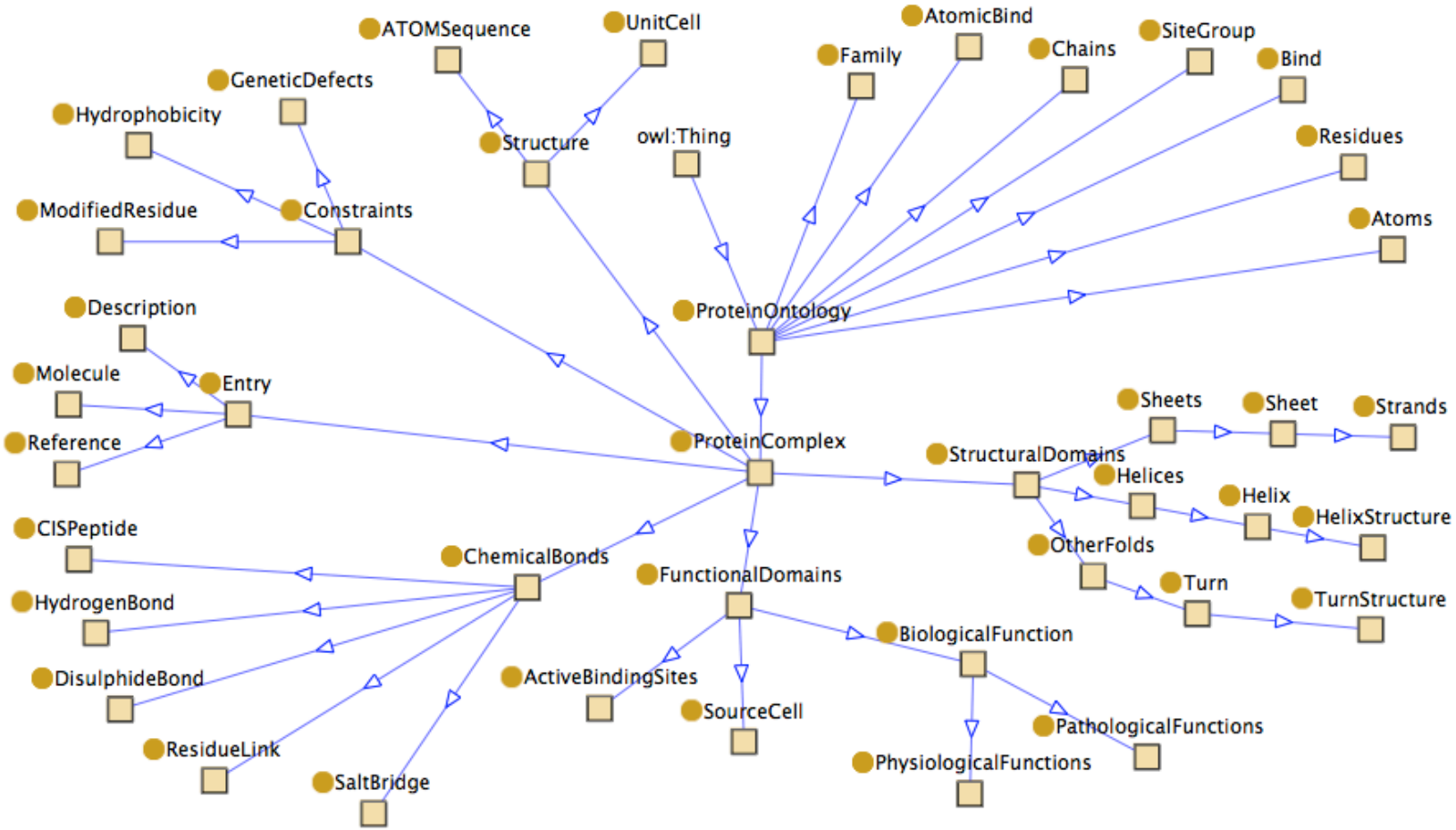

Figure 1: Protein Ontology Hierarchy Classification

4. SUMMARY

Nowadays many computational systems and databases have been developed to provide analysis and information on proteins. However, integration of the information is needed and so far this has not been possible as there was no common vocabulary available that could be used as a standard language. 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. It provides a unified controlled vocabulary both for annotation data types and for annotation data. It is accepted as part of Standardized Biomedical Ontologies at the National Center for Biomedical Ontologies (http://bioportal.bioontology.org/ontologies/3905) along with Gene Ontology and other biomedical ontologies.

REFERENCES

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BUTLER, H., CHERRY, J. C., CORRADI, J. & DOLINSKI, K. (2001) Creating the Gene Ontology Resource: Design and Implementation. *Genome Research,* 11**,** 1425-1433 [PMID: 11483584].

KING, O. D., FOULGER, R. E., DWIGHT, S., WHITE, J. & ROTH, F. P. (2003) Predicting gene function from patterns of annotation. *Genome Research,* 13**,** 896-904 [PMID: 12695322].

LEWIS, S. E. (2004) Gene Ontology: looking backwards and forwards. *Genome Biology,* 6**,** 103.1-103.4 [PMID: 15642104].

NATALE DA, ARIGHI CN, BARKER WC, BLAKE J, CHANG TC, HU Z, LIU H, SMITH B, WU CH. Framework for a protein ontology. BMC Bioinformatics. 2007, 8 Suppl 9:S1 [PMID: 18047702].

SAHOO SS, THOMAS C, SHETH A, YORK WS, TARTIR S: Knowledge modeling and its application in life sciences: a tale of two ontologies. In Proceedings of the 15th International Conference on World Wide Web Edited by: Carr L, De Roure D, Iyengar A, Goble CA, Dahlin M. New York, ACM Press; 2006, 317-326.

SIDHU, A. S., DILLON, T. S. & CHANG, E. (2007) Protein Ontology. IN CHEN, J. & SIDHU, A. S. (Eds.) *Biological Database Modeling.* New York, Artech House.

SIDHU, A. S., DILLON, T. S. & CHANG, E. (2005) An Ontology for Protein Data Models. In Zhang, Y.T., Roux, C. and Zhuang, T.G. (eds), 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2005). IEEE Engineering in Medicine and Biology Society, Shanghai, 6120-6123 [PMID: 17281660].

SIDHU, A. S., DILLON, T. S., SIDHU, B. S. & SETIAWAN, H. (2004) A Unified Representation of Protein Structure Databases. IN REDDY, M. S. & KHANNA, S. (Eds.) *Biotechnological Approaches for Sustainable Development.* India, Allied Publishers.

WHETZEL, P. L., PARKINSON, H., CAUSTON, H. C., FAN, L., FOSTEL, J., FRAGOSO, G., GAME, L., HEISKANEN, M., MORRISON, N., ROCCA-SERRA, P., SANSONE, S., TAYLOR, C., WHITE, J. & STOECKERT, C. J. (2006) The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics,* 22**,** 866-873 [PMID: 16428806].