An XML based semantic protein map

A. S. Sidhu, T. S. Dillon & H. Setiawan
Faculty of Information Technology, University of Technology, Sydney

Abstract

From the nature of the algorithms for data mining we note that an XML framework can be represented using graph matching algorithms. Various techniques currently exist for graph matching of data structures such as the Adjacency Matrix or Algebraic Representation of Graphs. The Graph Representation can be easily converted to a string representation. Both Graph and String Representations miss semantic relationships that exist in the data. These relationships can be captured by using semi-structured XML as a representation format. We already have an approach to integrate different data formats into a Unified Database. The technique is successfully applied to diverse Protein Databases in a Bioinformatics Domain. An XML representation of this comprehensive database preserving order and semantic relationships is already generated. In this paper we propose an approach to a Semantic Protein Map (PMAP) by building a shared ontology on our structured database model**.** This ontology can be used by various Bioinformatics researchers from one single site. This site will host mirrors of Protein Databases along with BIODB and have tools on Similarity Searching.

Keywords: bioinformatics, protein structures, biomedical ontologies, data integration, data semantics, semantic web.

1 Introduction

Bioinformatics is the field of science in which, biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. [1]. Data integration issues have stymied computer scientists and genetics alike for last 20 years, and yet successfully overcoming them is critical to success of genomics research as it transitions from wet-lab activity to an electronic-based

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

activity. This research is motivated by scientists striving to understand not only data that they have generated, but more importantly, the information implicit in these data, such as relationships between individual components. Only through this understanding will the scientists be able to successfully model and simulate entire genomes, cells, and ultimately entire organisms [2].

Many of the problems facing genomic data integration are related to data semantics – the meaning of data represented in a data source – and the differences between semantics within set of sources. The differences require addressing issues of concept identification, data transformation and concept overloading. These issues are addressed by identifying which abstract concepts are shared in each biological data source. Identification of shared concepts helps in locating conflicting information. Unfortunately, the semantics of biological data are usually hard to define precisely because they are not explicitly stated but are implicitly included in database design. Genomics or Proteomics (much less all of biology or life science) 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.

This is further complicated by observations that the specific community whose terminology is being used by data source is not explicitly identified and that the terminology evolves over time. For many of the larger, community data sources, the domain is oblivious – the Protein Data Bank (PDB) [3] handles protein structure information, the Swiss Prot [4] protein sequence database provides protein sequence information and useful annotations, etc. – but the terminology used may not reflect knowledge integration from multiple domains. The terminology used in smaller community data sources is typically selected based on usage model. Because these models can involve using concepts from different domains, data source will use whatever definitions are most intuitive, mixing the domains as needed, these are difficult to generalize. In this paper we will are proposing a Data Semantics Map for Protein Description, based on knowledge integrated from PDB [3], SwissProt [4] and OMIM [5]. The driving force for the proposed life science discovery model is turning complex, heterogeneous data into useful structured information and ultimately into systemized knowledge. The endeavour is simply the classic pathway for all science (Figure 1).

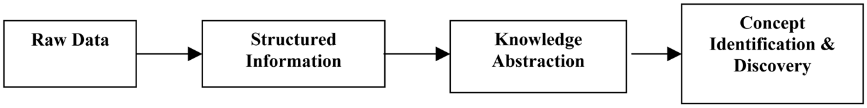

Figure 1: Classic science pathway.

We already have an approach to integrate different data formats into a Unified Database – BIOMAP [6]. The technique is successfully applied to diverse Protein Databases in Bioinformatics Domain. XML representation of this comprehensive database preserving order and semantic relationships is already generated. In this paper we propose an approach to for a Semantic Protein Map by building a shared ontology on this structured database. The Proposed Semantic Protein Map (PMAP) utilizes the power of XML to structure data and acquire knowledge abstractions form inherent structure of XML Data Files. XML based Protein Representation Framework provide the knowledge of semantic aspects of data, various integrated data sources and algorithms such as how to search, access, and retrieve information. The Semantic Protein Map also defines an Ontological Terminology to develop a loosely integrated knowledge integration systems mapping the understanding of various biological objects involved in Protein Engineering. Interpreting biological semantic relationships require understanding of biological meaning of data, beyond the layout of existing protein databases. This kind of information is provided by Unified Medical Language System (UMLS) [7].

2 Challenges in information integration for protein data sources

The scope of public protein data sources ranges from the comprehensive, multidisciplinary, community informatics center, supported by government public funds and sustained by team of specialised, to small data sources by individual investigators. The content of protein databases varies greatly, reflecting the broad disciplines and sub-disciplines across life sciences from proteomics and cell biology, to medicinal and clinical trails to ecology and biodiversity [2]. Data elements in public or proprietary protein databases are stored in heterogeneous data formats ranging from simple files to fully structured database systems that are often ad hoc, application specific and vendor specific. Scientific Literature, images and other free-text documents are commonly stored in unstructured or semi-structured formats (plain text files, HTML or XML files, binary files). Information Integration of protein data sources must consider the following characteristics [2]:

1. Diverse protein data are stored in autonomous data sources that are heterogeneous in data formats, data management systems, data schema and semantics.
2. Analysis of protein databases requires both database query activities and proper usage of computational analysis tools.
3. A Broad Spectrum of Knowledge Domains divides traditional Protein Domains in Molecular Biology.

Information Integration in Proteins faces challenges at technology level for data integration architectures and at semantic level for Meta – Data specifications, maintenance of data provenance and accuracy, ontology development for knowledge sharing and reuse, and Web representations for communication and collaboration. In this paper, the proposed Semantic Protein

Map addresses the following Information Integration Challenges for Protein Data – (1) Semantic Meta – Data integration of PDB [3], SwissProt [4] and OMIM[5]. (2) Knowledge Sharing & Reuse by using terminology from a shared ontology description, for Web Collaboration.

3 Existing data integration approaches

Over the past decade, enormous efforts and progress has been made in different data integration systems for biological databases in general. They can be roughly divided into three major categories according to access and architectures: the data warehousing approach, the distributed or federated approach, and the mediator approach. In this section we will briefly discuss these three approaches.

3.1 Data warehouse approach

The data warehouse approach assembles data sources into a centralized system with a global database schema and an indexing system for integration and navigation. These systems have proven very successful in health care. They require reliable operation and maintenance, and the underlying databases are under controlled environment, are fairly stable, and are structured. The biological data sources are very different from those contained in commercial databases. The biological databases are much more dynamic and unpredictable, and most of the public biological data sources use unstructured data formats. Given the sheer volume of data and broad range of biological databases, it would require substantial data warehouses encompassing diverse biological information such as sequence and structure and the various functions of biochemical pathways and genetic polymorphisms. Thus, limited data warehouses are popular solutions in life sciences for data mining of large databases [8].

3.2 The federation approach

The distributed or federated integration approaches do not require a centralized persistent database, and thus the underlying data sources remain autonomous. The federated systems maintain a common data model and rely on schema mapping to translate heterogeneous source database schema into target schema for integration. A data dictionary is used to manage various schema components. These systems typically rely on interfaces such as Common Object Request Broker Architecture (CORBA), an open standard by Object Management Group (OMG) to facilitate interoperation of disparate components [9, 10]. The mmCIF Data Format [11] of Protein Data Bank is a perfect example. Its Macromolecular Specification [12] is developed by International Union of Crystallography (IUCr) and OMG. It is based on Dictionary Definition Language, that in broader sense is quite close to an Ontology In biological databases arena, schema changes in data sources are frequent; the maintenance of a common data dictionary could be costly in large federated systems.

3.3 The mediator approach

The most flexible data integration adopt a mediator approach that introduces an intermediate processing layer to decouple the underlying heterogeneous distributed data sources and the client layer of end users and applications. The mediator layer is a collection of software components performing the task of data integration. Most mediator systems use a wrappers layer to handle tasks of access, retrieval and translation. The mediator layer performs the core function of data transformation and integration and communicates with wrappers and user application layers. The integration system provides an internal common data model for abstraction of incoming data derived from heterogeneous data sources. This approach is most suitable for specific investigations that need to access most up-to-date data. The Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) [13, 14] provides an excellent example of mediator approach using a global ontology to facilitate queries over multiple data sources.

4 Semantic protein meta – data specification

R. Karp [15] once said "There are many protein databases out there, and each one chooses to conceptualise or represent proteins in its schema in a different way. So someone who wants to issue a query using 10 protein databases has to examine each database to figure out how it encodes proteins, what information it encodes, what field name it uses and what units of measurement it uses"

Recently biological IT community has been picking up the momentum to adopt the merging XML technology for biological Web services for exchange of data. Here we propose a XML based Meta – Data Specification for Protein Structure Data Representation. The proposed XML Meta – Data Specification addresses the following issues:

1. Semi Structured XML handles unstructured data. XML database technologies do not have regular structure making them suitable to hold unstructured data in table form.
2. There is no dingle biological mapping to which Protein Data can refer. Individual users of data will have their own prospectives. The proposed semantic mapping is an attempt to address the problem.
3. The Proposed Protein Meta – Data Specification is a way of enabling distributed data to be accessed in a form whereby the semantics of data are explicit. This assists in easy easier data extraction and knowledge acquisition by computer programs.

5 Semantic protein map outline

Peptides and proteins are constructed from sequences of amino acids. Protein biosynthesis is a very complex process, represented by the central paradigm (Figure 2).

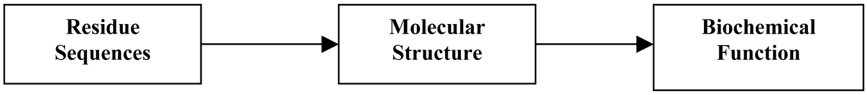

Figure 2: Central paradigm of proteomics.

The proposed Semantic Protein Map Description takes into consideration the following key principles of protein biosynthesis:

Protein Domains – Protein Domain is defined as a chain in a protein structure that can fold independently into a stable three dimensional structure.

Secondary Structure Packing – Most protein structures contain a significant amount of secondary structure (α – helices, β – strands or coil). Three dimensional structures can be determined by deducing which stretches of amino acid sequences should adopt each type of secondary structure, and then how these secondary structure elements are packed together.

Threading – Threading, more commonly, known as fold recognition is a method that can be used to suggest general structure for a new protein. The concept is to 'thread' the unknown sequence through a set of sequences of known three-dimensional proteins, typically chosen to represent common structural classes.

Protein Folding – Proteins typically adopt a single structure, corresponding to global minimum of free energy under physiological conditions. Proteins generally fold into this unique state in just a few seconds from any starting conformation. Proteins can be unfolded using high temperatures and non aqueous solutions. However Protein Folding is reversible.

External Effects on Protein Structure – Factors of Mutation and Chemical Environment under physiological conditions are important in deciding stable conformation of Protein that exists in a cell.

6 Terminologies for semantic protein map

Models represent **Aspects**, which denote a coherent set of properties of Protein Structure like Atom, Residue, Chain, Protein Domains etc. Aspects anchor the Protein Map in Real World. Experimental biologists make **Observations** about the Protein Engineering phenomenon. Protein Map once instantiated yield **Interpretation** though analysis. Biologists deduce **Hypothesis** from these Interpretations. Data Mining Algorithms derive Associations from the Interpretations. **Constraints** like Genetic Defects and Environmental Effects that affect the final conformation of Protein Structure. **Context** represents the Data Semantics inherent in Protein Structure Description.

7 Meta model of semantic protein map

To map out Protein Structure Data Space more systematically, and to identify challenges of mapping Sequence, Structure, and Function of Protein we use information model shown in Figure 3.

The Primary Goal of Semantic Protein Mapping (PMAP) is to define a shared, structured and context based vocabulary to annotate molecular attributes across various Protein Structures. The Overall Structure of PMAP is shown in Figure 4.

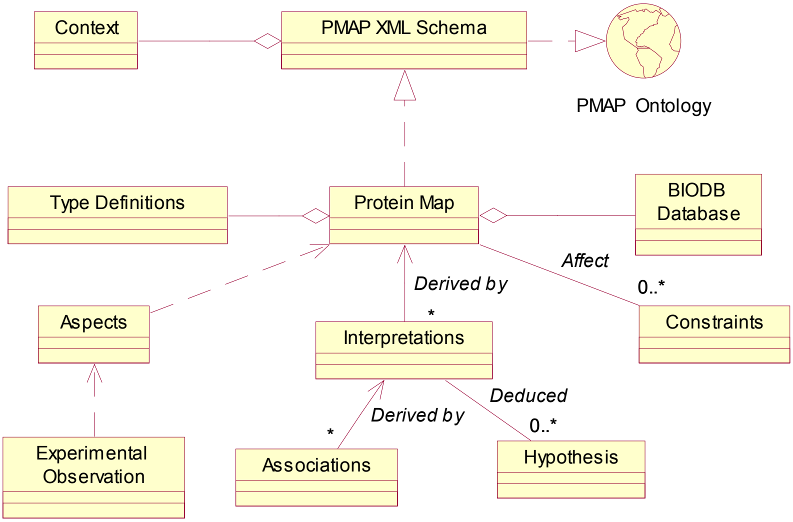

Figure 3: PMAP meta model.

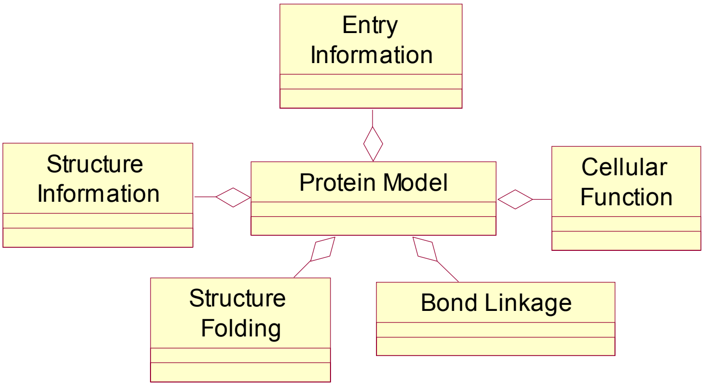

Figure 4: PMAP structure.

The Semantic Protein Map (PMAP) describes various aspects of Protein Engineering by categorizing information about a Protein Structure into following:

7.1 Entry information

Type Definition of Entry (as shown in Figure 5) describes – (1) General Protein Information (InfoType), (2) Information about Compounds present in Protein

(CompoundType) and (3) Information about Citation of Protein Structures in Literature (CitationType).

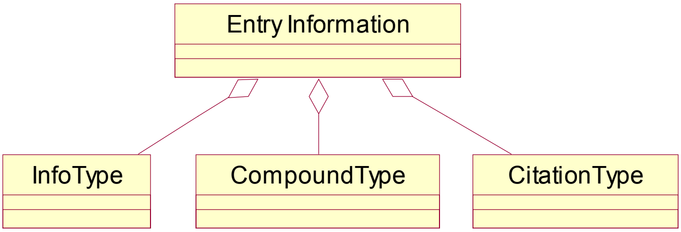

Figure 5: Entry type definition.

7.2 Structure information

Type Definition of Structure (as shown in Figure 6) describes – (1) Protein Sequence & Structure information using concept of "ATOM Sequences" (ATOMSequenceType) and (2) Unit Cell Information (UnitCellType).

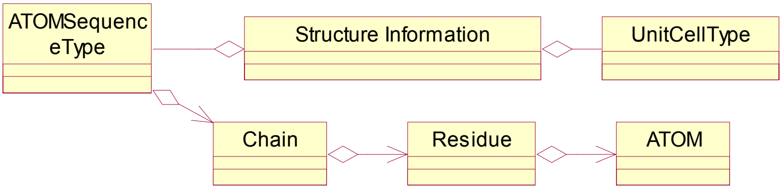

Figure 6: Structure type definition.

7.3 Cellular Function information

Type Definition of Cellular Function (as shown in Figure 7) describes – (1) Information about various Protein Domains known (DomainType), and (2) Information about Source Cell and the Cellular Environment in which Protein Exists.

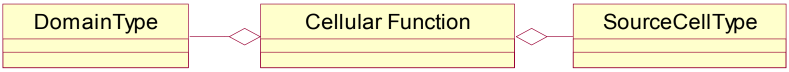

Figure 7: Cellular function definition.

7.4 Structure Folding information

Type Definition of Structure Folding (as shown in Figure 8) describes Protein Folding Information of Secondary Structure of Protein as: (1) HelixType, (2) SheetType and (3) TurnType.

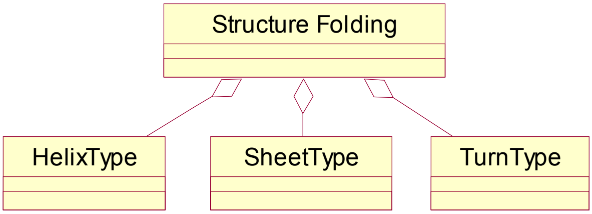

Figure 8: Structure folding definition.

7.5 Bond Linkage information

Type Definition of Bond Linkage Information (as shown in Figure 9) describes Bond Links of Protein as: (1) HydrogenBondType, (2) DisulphideBondType, (3) ResidueLinkType, (4) SiteType, (5) CISPeptideType and (6) SaltBridgeType.

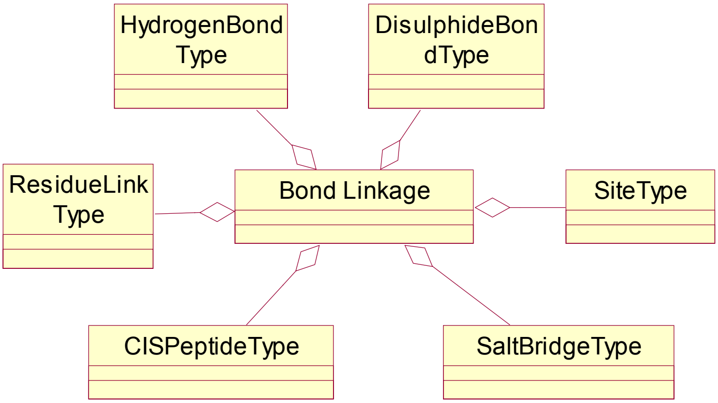

Figure 9: Bond linkage definition.

8 Conclusion

BIODB's [6] XML based Protein Model structures heterogeneous protein data sources. PMAP Ontology defines Data Semantics on top of BIODB [6], thus providing a single unified ontology that is able to identify accurately the knowledge contained in all protein data sources. We establish a correspondence through Protein Map, an explicit formal specification of how to represent objects, concepts and relationships that hold for them in a Protein Structure Representation. We can link our Protein Map with Gene Ontology [16] and RiboWEB [17] to dependably diagnose health issues and identify novel treatments. Having this Global Protein Map would allow mappings between related concepts to be easily identified by Association Rule Mining and Classification Algorithms in Data Mining.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

References

[1] NCBI (2002). Just the Facts: A Basic Introduction to the science underlying NCBI Resources. A Science Primer. November 2002.

[2] Lacroix, Z. and T. Critchlow (2003). Bioinformatics Managing Scientific Data. San Francisco, Morgan Kaufmann Publishers (Elsevier Science).

[3] M. Berman, H., et al., The Protein Data Bank. Nucleic Acids Research, 2000. 28(1): p. 235-242.

[4] Bairoch, A. and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Research, 1997. 25(1): p. 31-36.

[5] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000.

[6] Sidhu, A. S., T. S. Dillon, et al. (2004). Comprehensive Protein Database Representation. Eighth Annual Conference on Research in Computational Molecular Biology (RECOMB 2004), San Diego, CA, USA.

[7] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 Aug;32(4):281-91.

[8] R. Resnick (2000). "Simplified Data Mining." Drug Discovery and Development(October 2000): 51-52.

[9] K. Jungfer, G. Cameron, et al. (1999). EBI: CORBA and EBI Databases. Bioinformatics: Databases and Systems. S. Letovsky. Norwell, MA, Kluwer Academic Publishers: 245-254.

[10] Siepel, A. C., A. N. Tolopko, et al. (2001). "An Integration Platform for Heterogeneous Bioinformatics Software Components." IBM Systems Journal 40(2): 570-591.

[11] Westbrook, J. D. and P. E. Bourne (2000). "STAR/mmCIF: An Ontology for molecular structure." Bioinformatics 16(2): 159-168.

[12] OMG (2002). Macromolecular Structure Specification. Version 1.0.

[13] Patton, N. W., R. Stevens, et al. (1999). Query Processing in TAMBIS Bioinformatics Source Integration System. 11th IEEE International Conference on Scientific and Statistical Database Management.

[14] R. Stevens, P. Baker, et al. (2000). "TAMBIS : Transparent Access to Multiple Bioinformatics Information Sources." Bioinformatics 16(2): 184-186.

[15] Richard M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, Complexity of Computer Computations, pages 85–104. Plenum Press, New York, 1972.

[16] The Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32: D258-D261.

[17] Chen, R. O., R. Felciano, et al. (1997). RIBOWEB: linking structural computations to a knowledge base of published experimental data. International Conference of Intelligent Systems in Molecular Biology.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9