Biological Database Modeling

For a listing of recent titles in the
Artech House Bioinformatics & Biomedical Imaging Series,
turn to the back of this book.

Biological Database Modeling

Jake Chen
Amandeep S. Sidhu

Editors

ARTECH
HOUSE

BOSTON | LONDON
artechhouse.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the U.S. Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN-13: 978-1-59693-258-6

Cover design by Igor Valdman

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Database management systems (DBMS) are designed to manage large and complex data sets. In the past several decades, advances in computing hardware and software and the need to handle rapidly accumulating data archived in digital media have led to significant progress in DBMS research and development. DBMS have grown from simple software programs that handled flat files on mainframe computers, which were prohibitively expensive to all but a few prestigious institutions, into today's popular form of specialized software platforms underpinning wide ranges of tasks, which include business transactions, Web searches, inventory management, financial forecasts, multimedia development, mobile networks, pervasive computing, and scientific knowledge discovery. Technologies of DBMS have also become increasingly sophisticated, diverging from generic relational DBMS into object-relational DBMS, object-oriented DBMS, in-memory DBMS, semantic Webs data store, and specialized scientific DBMS. Given the sustained exponential data growth rate brought forth by continued adoption of computing in major industries and new inventions of personal digital devices, one can safely predict that DBMS development will continue to thrive in the next millennium.

In this book, we want to share with our readers some fresh research perspectives of post-genome biology data management, a fast-growing area at the intersection of life sciences and scientific DBMS domains. Efficient experimental techniques, primarily DNA sequencing, microarrays, protein mass spectrometers, and nanotechnology instruments, have been riding the wave of the digital revolution in the recent 20 years, leading to an influx of high-throughput biological data. This information overload in biology has created new post-genome biology studies such as genomics, functional genomics, proteomics, and metabolomics—collectively known as "omics" sciences in biology. While most experimental biologists are still making the transition from one-gene-at-a-time type of studies to the high-throughput data analysis mindset, many leaders of the field have already begun exploring new research and industrial application opportunities. For example, managing and interpreting massive omics data prelude ultimate systems biology studies, in which one may analyze disparate forms of biological data and uncover coordinated functions of the underlying biological systems at the molecular and cellular signalling network level. On the practical side, understanding diverse intricate interplays between environmental stimuli and genetic predisposition through omics evidence can help pharmaceutical scientists design drugs that target human proteins with high therapeutic values and low toxicological profiles. With data management tools to handle terabytes of omics data already released in the public domain, the promise of post-genome biology looms large.

Compared with data from general business application domains, omics data has many unique characteristics that make them challenging to manage. Examples of these data management challenges are:

1. Omics data tends to have more complex and more fast-evolving data structures than business data. Biological data representation often depends on scientific application scenarios. For example, biological sequences such as DNA and proteins can be either represented as simple character strings or connected nodes in three-dimensional spatial vectors. Data representation is an essential first step.
2. Omics data is more likely to come from more heterogeneously distributed locations than business data. To study systems biology, a bioinformatics researcher may routinely download genome data from the Genome Database Center at the University of California, Santa Cruz, collect literature abstracts from the PubMed database at the National Library of Medicine in Maryland, collect proteome information from the Swiss-Prot database in Switzerland, and collect pathway data from the KEGG database in Japan. Data integration has to be carefully planned and executed.
3. Omics data tends to reflect the general features of scientific experimental data: high-volume, noisy, formatted inconsistently, incomplete, and often semantically incompatible with one another. In contrast, data collected from business transactions tends to contain far fewer errors, is often more accurate, and shows more consistencies in data formats/coverage. Meticulous data preprocessing before knowledge discovery are required.
4. Omics data also lags behind business data in standard development. For example, Gene Ontology (GO) as a standard to control vocabularies for genes was not around until a decade ago, whereas standards such as industrial product categories have been around for decades. The ontology standards and naming standards for pathway biology are still under development. This makes it difficult to perform mega collaboration, in which cross-validation of results and knowledge sharing are both essential.

Despite all the challenges, modeling and managing biological data represent significant discovery opportunities in the next several decades. The human genome data bears the ultimate solutions of expanding the several thousand traditional molecular drug targets into tens of thousands genome drug targets; molecular profiling information, based on individuals using either the microarrays or the proteomics platform, promises new types of molecular diagnostics and personalized medicine. As new applications of massive biological data emerge, there will be an increasing need to address data management research issues in biology.

In this compiled volume, we present to our readers a comprehensive view of how to model the structure and semantics of biological data from public literature databases, high-throughput genomics, gene expression profiling, proteomics, and chemical compound screening projects. The idea of compiling this book, which we found to be unique, stems from the editors' past independent work in bioinformatics and biological data management. While topics in this area are diverse and interdisciplinary, we focused on a theme for this book—that is, how to model and manage omics biological data in databases. By promoting this theme for the past decade among ourselves and the contributing authors of this book, we have contributed to solving complex biological problems and taking biological database management problems to the next level. We hope our readers can extract similar insights by using this book as a reference for future related activities.

There are 11 chapters presented in this book. Individual chapters have been written by selected accomplished research teams active in the research of respective topics. Each chapter covers an important aspect of the fast-growing topic of biological database modeling concepts. Each chapter also addresses its topic with varying degrees of balance between computational data modeling theories and real-world applications.

In Chapters 1 through 5, we introduce basic biological database concepts and general data representation practices essential to post-genome biology. First, biological data management concepts are introduced (Chapter 1) and major public database efforts in omics and systems biology studies are summarized (Chapter 2). Then, biomedical data modeling techniques are introduced (Chapter 3). Next, Gene Ontology as an established basic set of controlled vocabulary in genome database annotations is described (Chapter 4). Finally, the latest research on protein ontology and the use of related semantic webs technologies are presented to enable readers to make the connection between emerging biological data collection and integration trends (Chapter 5).

In Chapters 6 through 9, we examine in detail how to develop data management techniques to process and analyze high-throughput biological data through case studies. First, quality control techniques to reduce variations during experimental data collection steps are described (Chapter 6). Then, biological sequence management experience for a fungi genomics project is discussed (Chapter 7). Next, data management and data integration methods for microarray-based functional genomics studies are investigated (Chapter 8). Finally, data management challenges and opportunities for mass spectrometry based expression proteomics are presented (Chapter 9).

In Chapters 10 and 11, we delve into the practical aspect, demonstrating how to apply biological data management for drug discoveries. First, fundamental drug discovery concepts based on macromolecular structural modeling are introduced (Chapter 10); then, a data management software system that implements high-throughput drug compound screenings is discussed (Chapter 11) to conclude the book.

We hope this book will become a useful resource for bioinformatics graduate students, researchers, and practitioners interested in managing post-genome biological data. By studying the techniques and software applications described in this book, we hope that bioinformatics students will use the book material as a guide to acquire basic concepts and theories of post-genome biological data management, bioinformatics practitioners will find valuable lessons for building future similar biological data management systems, and researchers will find rewarding research data management questions to address in the years to come.

Acknowledgments

We wish to thank all of the authors for sharing their insightful knowledge and making excellent contributions to this book based on their active research in biological data management. This book would not have been completed without the tremendous efforts, held to the highest standards, of all the authors, each of whom spent numerous hours over many drafts in preparing, collating, and revising their writings over the past 2 years. During the publishing process, many colleagues also helped and they deserve our whole-hearted appreciation. They are: David Wong from Indiana University, who provided legal advice for the contract agreement; Dr. Zongmin Ma from Northwestern University of China, who provided assistance in the initial conceptualization and execution of the book publishing process; Susan Lagerstrom-Fife from Springer Science+Business Media, Inc., who guided us from a publisher's perspective while we explored various publication options; Wayne Yuhasz from Artech House Publishers, whose persistence and dedication to timely assistance finally won us over in making our book part of the Artech House Bioinformatics & Biomedical Imaging Series; and Barbara Lovenvirth from Artech House Publishers, who assisted us throughout the final publication process.

We also want to thank the general support from our home institutions: Indiana University School of Informatics at Indianapolis, Indiana; Purdue University School of Science, Department of Computer and Information Sciences at Indianapolis, Indiana; and Curtin University of Technology, Perth, Australia.

Last, but not least, Jake Chen wishes to thank his family, including his wife, Mabel Liu, for her unbounded love and support in assuming additional responsibilities so that the book project could be completed.

Jake Chen
Indianapolis, Indiana
Amandeep S. Sidhu
Perth, Australia
Editors
October 2007

CHAPTER 5

Protein Ontology

Amandeep S. Sidhu, Tharam S. Dillon, and Elizabeth Chang

Two factors dominate current developments in structural bioinformatics, especially in protein informatics and related areas: (1) the amount of raw data is increasing, very rapidly; and (2) successful application of data to biomedical research requires carefully and continuously curated and accurately annotated protein databanks. In this chapter, we introduce the concepts for annotating protein data using our protein ontology. We first describe and review existing approaches for protein annotation. We then describe the advantages of semantic integration of protein data using Web ontology language, in comparison to annotating using automatic search and analyzing using text mining. The rest of chapter is devoted to the use of the protein ontology for the annotation of protein databases. The protein ontology is available at http://www.proteinontology.info/.

5.1 Introduction

A large number of diverse bioinformatics sources are available today. The future of biological sciences promises more data. No individual data source will provide us with answers to all the queries that we need to ask. Instead, knowledge has to be composed from multiple data sources to answer the queries. Even though multiple databases may cover the same data, their focus might be different. For example, even though Swiss-Prot [1–3] and PDB [4–7] are both protein databases, we might want to get information about sequence as well as structure of a particular protein. In order to answer that query, we need to get data about the protein from both sources and combine them in a consistent fashion [8]. In this postgenomic era, isolating specific data from heterogeneous protein data sources has become a major issue. Extracting relevant protein data without omitting data and without introducing irrelevant information is a challenge. The main problems lie in interpreting protein nomenclature and the multitude of synonyms and acronyms, which can be used to describe a single protein, identifying data provenance, and extracting data from computationally unprocessed natural language.

5.2 What Is Protein Annotation?

As biological databases undergo a very rapid growth, two major consequences are emerging. First, an efficient utilization of databases to provide easy management and access to this data is continuously developing. A second consequence is a necessary formalization of biological concepts and relationships among these concepts via the creation of ontologies.

In the context of protein data, annotation generally refers to all information about a protein other than protein sequence. In a collection of protein data, each protein is labeled at least by an identifier and is usually complemented by annotations as free text or as codified information, such as names of authors responsible for that protein, submission date of protein data, and so on. Annotations become a challenge in proteomics considering the size and complexity of protein complexes and their structures.

For our purposes, we will mainly deal with two main sources of protein annotations: (1) those taken from various protein data sources submitted by the authors of protein data themselves from their published experimental results; and (2) those that we name annotation that are obtained by an annotator or group of annotators through analysis of raw data (typically a protein sequence or atomic structure description) with various tools that extract biological information from other protein data collections.

5.3 Underlying Issues with Protein Annotation

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, structural similarity, or functional identification. The success of this method is dependent on the family the protein belongs to. Some proteins have a high degree of sequence identity, structural similarity, or similarity in functions that are unique to members of that family. Consequently, this approach cannot be generalized to integrate the protein data. Clearly, these traditional approaches have limitations in capturing and integrating data for protein annotation. For these reasons, we have adopted an alternative method that does not rely on keywords or similarity metrics, but instead uses ontology. Briefly, *ontology* is a means of formalizing knowledge; at the minimum, ontology must include concepts or terms relevant to the domain, definitions of concepts, and defined relationships between the concepts.

We have built protein ontology [9–17] to integrate protein data formats and provide a structured and unified vocabulary to represent protein synthesis concepts. 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, interaction between two protein molecules, and interactions between protein and other macromolecules at cellular level. PO also helps to codify proteomics data for analysis by researchers. Before we discuss the PO framework in detail, in the next section we provide an overview of various protein databanks and earlier attempts to integrate protein data from these data sources.

5.3.1 Other Biomedical Ontologies

In this section we will discuss various biomedical ontology works related to protein ontology. Gene ontology (GO) [18, 19] defines a hierarchy of terms related to genome annotation. GO is a structured network consisting of defined terms and relationships that describe molecular functions, biological processes, and cellular components of genes. GO is clearly defined and modeled for numerous other biological ontology projects. So far, GO has been used to describe the genes of several model organisms (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Mus musculus*, and others).

RiboWEB [20] is an online data resource for Ribosome, a vital cellular apparatus. It contains a large knowledge base of relevant published data and computational modules that can process this data to test hypotheses about ribosome's structure. The system is built around the concept of ontology. Diverse types of data taken principally from published journal articles are represented using a set of templates in the knowledge base, and the data is linked to each other with numerous connections.

Protein Data Bank (PDB) has recently released versions of the PDB Exchange Dictionary and the PDB archival files in XML format, collectively named PDBML [21]. The representation of PDB data in XML builds from content of the PDB Exchange Dictionary, both for assignment of data item names and defining data organization. PDB exchange and XML representations use the same logical data organization. A side effect of maintaining a logical correspondence with PDB exchange representation is that PDBML lacks the hierarchical structure characteristic of XML data.

PRONTO [22] is a directed acyclic graph (DAG)-based ontology induction tool that constructs a protein ontology including protein names found in MEDLINE abstracts and in UniProt. It is a typical example of text mining the literature and the data sources. It cannot be classified as protein ontology as it only represents relationship between protein literatures and does not formalize knowledge about protein synthesis process. Ontology for protein domain must contain terms or concepts relevant to protein synthesis, describing protein sequence, structure, and function and relationships between them. While defining PO we made an effort to emulate the protein synthesis and describe the concepts and relationships that describe it.

There is a need for an agreed-upon standard to semantically describe protein data. PO addresses this issue by providing clear and unambiguous definitions of all major biological concepts of protein synthesis process and the relationships between them. PO provides a unified controlled vocabulary both for annotation data types and for annotation data itself.

5.3.2 Protein Data Frameworks

The identification of all the genes that encode proteins in the genome of an organism is essential but not sufficient for understanding how these proteins function in making up a living cell. The number of different fundamental proteins in an organism often substantially exceeds the number of genes due to generation of protein isoforms by alternative RNA processing as well as by covalent modifications of precursor proteins. To cope with the complexity of protein sequence and functional information, annotated databases of protein sequence, structure, and function with high interoperability are needed. The major protein databases are the most comprehensive sources of information on proteins. In addition to these universal databases that cover proteins from all the species, there are collections that store information about specific families or groups of proteins, or about proteins of specific organisms. Here we will give a brief overview of major protein collections and their corresponding annotations.

5.3.2.1 Worldwide PDB (wwPDB)

The wwPDB [23] represents a milestone in the evolution of PDB, which was established in 1971 at Brookhaven National Laboratory as the sole international repository for three-dimensional structural data of biological macromolecules. Since July 1, 1999, the PDB has been managed by three member institutions of the RCSB: Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology. The wwPDB recognizes the importance of providing equal access to the database both in terms of depositing and retrieving data from different regions of the world. Therefore, the wwPDB members will continue to serve as deposition, data processing, and distribution sites. To ensure the consistency of PDB data, all entries are validated and annotated following a common set of criteria. All processed data is sent to the RCSB, which distributes the data worldwide. All format documentation will be kept publicly available and the distribution sites will mirror the PDB archive using identical contents and subdirectory structure. However, each member of the wwPDB will be able to develop its own Web site, with a unique view of the primary data, providing a variety of tools and resources for the global community.

5.3.2.2 Universal Protein Resource (UniProt)

UniProt [1] joins three major protein databases: PIR-PSD [24], Swiss-Prot [3], and TrEMBL [2]. The Protein Information Resource (PIR) provides an integrated public resource of protein informatics. PIR produces the Protein Sequence Database (PSD) of functionally annotated protein sequences. Swiss-Prot is a protein knowledge base established in 1986 and maintained collaboratively by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). It strives to provide a high level of annotation, a minimal level of redundancy, a high level of integration with other biomolecular databases, and extensive external documentation. The translation of EMBL nucleotide sequence database (TrEMBL), a supplement to Swiss-Prot, was created in 1996. This supplement contains computer annotated protein sequences not yet integrated with Swiss-Prot. Together, PIR, EBI, and SIB maintain and provide UniProt, a stable, comprehensive, fully classified, and accurately annotated protein knowledge base.

5.3.2.3 Classification of Protein Families

Classification of proteins provides valuable clues of structure, activity, and metabolic role. A number of different classification systems have been developed in recent years to organize proteins. The various existing classification schemes include: (1) hierarchical families of proteins, such as the superfamilies or families in the PIR-PSD, and protein groups in ProntoNet [25]; (2) protein family domains such as those in Pfam [26] and ProDom [27]; (3) sequence motifs or conserved regions as in PROSITE [28] and PRINTS [29]; (4) structural classes, such as in SCOP [30] and CATH [31]; and (5) integrations of various family classifications such as iProClass [32] and InterPro [33]. While each of these databases is useful for particular needs, no classification scheme is by itself adequate for addressing all protein annotation needs.

InterPro is an integrated resource of PROSITE, PRINTS, Pfam, ProDom, SMART [34], and TIGRFAMs [35] for protein families, domains, and functional sites. Each entry in InterPro includes a unique name and short name; an abstract, which provides annotation about protein matching the entry; literature references and links back to relevant databases; and a list of precomputed matches against the whole of SwissProt and TrEMBL.

PIR defines closely related proteins as having at least 50% sequence identity; such sequences are automatically assigned to the same family. The families produced by automatic clustering can be refined to make groups that make biological sense. A PIR superfamily is a collection of families. Sequences in different families in same superfamily have as little as 15–20% sequence identity. The PIR superfamily / family concept [36] is the earliest protein classification based on sequence similarity, and is unique in providing nonoverlapping clustering of protein sequences into a hierarchical order to reflect evolutionary relationships.

5.3.2.4 Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders. OMIM focuses primarily on inherited or heritable, genetic diseases. It is also considered to be a phenotypic companion to the human genome project. OMIM is based upon the text *Mendelian Inheritance in Man* [37], authored and edited by Victor A. McKusick and a team of science writers and editors at Johns Hopkins University and elsewhere. *Mendelian Inheritance in Man* is now in its twelfth edition. The database contains textual information and references. OMIM is primarily used by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine.

5.3.3 Critical Analysis of Protein Data Frameworks

Semantics of protein data are 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. This is further complicated by observations that the specific community whose terminology is being used by a 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 PDB handles protein structure information, the Swiss-Prot protein sequence database provides protein sequence information and useful annotations, and so on. The terminology used in these major data banks does not reflect knowledge integration from multiple protein families. The wwPDB is an effort to integrate protein data from PDB and other major protein databases at EBI, but the work is at too early a stage at the moment to comment. Furthermore, in smaller community data sources terminology is typically selected based on functionality of data source or usage model. Consequently, as queries to protein data models discussed in this section can involve using concepts from other data sources, the data source answering the query will use whatever definitions are most intuitive, annotating the knowledge from various protein families as needed. This kind of protein annotation will be difficult to generalize for all kinds of proteins.

5.4 Developing Protein Ontology

One of the major motivations for developing protein ontology was introduction of the Genomes to Life Initiative [38, 39] close to the completion of Human Genome Project (HGP), which finished in April 2003 [40]. Lessons learned from HGP will guide ongoing management and coordination of GTL. The overall objective of the Protein Ontology Project is: “To correlate information about multiprotein machines with data in major protein databases to better understand sequence, structure and function of protein machines.” The objective is achieved to some extent by creating databases of major protein families, based on the vocabulary of PO.

As technologies mature, the shift from single annotation databases being queried by Web-based scripts generating HTML pages to annotation repositories capable of exporting selected data in XML format, either to be further analyzed by remote applications or to undergo a transformation stage to be presented to user in a Web browser, will undoubtedly be one of the major evolutions of protein annotation process. XML is a markup language much like HTML, but XML describes data using hierarchy. An XML document uses the schema to describe data and is designed to be self-descriptive. This allows easy and powerful manipulation of data in XML documents. XML provides syntax for structured documents, but imposes no semantic constraints on the meaning of these documents.

Resource Description Framework (RDF) is a data model for objects or resources and relations between them; it provides a simple semantics for this data model, and these data models can be represented in XML syntax. RDF Schema is a vocabulary for describing properties and classes of RDF resources, with semantics for generalization hierarchies of such properties and classes.

To efficiently represent the protein annotation framework and to integrate all the existing data representations into a standardized protein data specification for the bioinformatics community, the protein ontology needs to be represented in a format that not only enforces semantic constraints on protein data, but can also facilitate reasoning tasks on protein data using semantic query algebra. This motivates the representation of the Protein Ontology Model in Web Ontology Language (OWL). OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing additional vocabulary along with a formal semantics. OWL provides a language for capturing declarative knowledge about protein domain and a classifier that allows reasoning about protein data. Knowledge captured from protein data using OWL is classified in a rich hierarchy of concepts and their interrelationships. OWL is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval, and querying. We investigated the use of OWL for making PO using the Protégé OWL Plug-in. OWL is flexible and powerful enough to capture and classify biological concepts of proteins in a consistent and principled fashion. OWL is used to construct PO that can be used for making inferences from proteomics data using defined semantic query algebra.

5.5 Protein Ontology Framework

The ultimate goal of protein annotator framework or PO is to deduce from proteomics data all its biological features and describe all intermediate structures: primary amino acid sequence, secondary structure folds and domains, tertiary three-dimensional atomic structure, quaternary active functional sites, and so on. Thus, complete protein annotation for all types of proteins for an organism is a very complex process that requires, in addition to extracting data from various protein databases, the integration of additional information such as results of protein experiments, analysis of bioinformatics tools, and biological knowledge accumulated over the years. This constitutes a huge mass of heterogeneous protein data sources that need to rightly represented and stored. Protein annotators must be able to readily retrieve and consult this data. Therefore protein databases and man-machine interfaces are very important when defining a protein annotation using protein ontology.

The process of development of a protein annotation based on our protein ontology requires a significant 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), and the lack of standardization in nomenclature.

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 communities agree upon concepts underlying the data. PO provides a framework of structured vocabularies and a standardized description of protein concepts which help to achieve this agreement and achieve uniformity in protein data representation.

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, 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. PO provides the concepts necessary to describe individual proteins, but it does not contain individual protein instances. The PO Instance Store contains individual instances for protein complex in the Web Onthology Language (OWL) format.

5.5.1 The ProteinOntology Concept

The main concept in PO is ProteinOntology. For each instance of protein that is entered into PO, the submission information is entered for the ProteinOntology Concept. The ProteinOntology Concept has the following attributes: ProteinOntologyID and ProteinOntologyDescription. The ProteinOntologyID has the following format: PO000000029.

5.5.2 Generic Concepts in Protein Ontology

There are seven subconcepts of the ProteinOntology Concept, called generic concepts, which are used to define complex concepts in other PO classes: Residue, Chain, Atom, Family, AtomicBind, Bind, and SiteGroup. These generic concepts are reused for other definitions of complex concepts in PO. Details and properties of residues in a protein sequence are defined by instances of Residue Concept. Instances of chains of residues are defined in Chain Concept. Three-dimensional structure data of protein atoms is represented as instances of Atom Concept. Defining chain, residue, and atom as individual concepts has the benefit that any special properties or changes affecting a particular chain, residue, or atom can be easily added. Family Concept represents protein superfamily and family details of proteins. Data about binding atoms in chemical bonds like a hydrogen bond, residue links, and salt bridges is entered into the 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 the ontology as an instance of Bind Concept. All data related to site groups of the active binding sites of proteins are defined as instances of SiteGroup Concept. Representation of instances of residues and chains of residues are shown as follows:

```
<Residues>
<Residue>LEU</Residue>
<ResidueName>LEUCINE</ResidueName>
<ResidueProperty>1-LETTER CODE: L; FORMULA: C6 H13 N1 O2;
```

```
MOLECULAR WEIGHT: 131.17</ResidueProperty>
</Residues>

<Chains>
<Chain>D</Chain>
<ChainName>CHAIN D</ChainName>
</Chains>
```

5.5.3 The ProteinComplex Concept

The root concept for the definition of protein complexes in the protein ontology is ProteinComplex. The ProteinComplex Concept defines one or more proteins in the complex molecule. There are six subconcepts of ProteinComplex: Entry, Structure, StructuralDomains, FunctionalDomains, ChemicalBonds, and Constraints. These subconcepts define sequence, structure, function, and chemical bindings of the protein complex.

5.5.4 Entry Concept

Entry Concept specifies the details of a protein or a protein complex which are entered into the knowledge base of protein ontology. Protein entry details are entered into Entry Concept as instances of EntrySuperFamily, EntryFamily, SourceDatabaseID, SourceDatabaseName, SubmissionDate, and Classification. These attributes describe the entry in the original protein data source from where it was taken. Entry has three subconcepts: Description, Molecule, and Reference. The Description subconcept describes data about the title of the entry, the authors of the entry, the experiment that produced the entry, and the keywords describing the entry. The second subconcept of Entry is Molecule, which is simply any chemically distinct molecule or compound in a protein complex. MoleculeID uniquely identifies a molecule. MoleculeName is the chemical name of the molecule. MoleculeChain refers to the chain description. BiologicalUnit instance describes the larger biological unit of which the molecule is a part. Engineered identifies whether the molecule is engineered using recombinant technology or chemical synthesis. A specific domain or region of the molecule is defined using Fragment. Mutated molecules of the protein have Mutations Information. Details about various mutations are described in the GeneticDefects Class. A list of synonyms for molecule name are in Synonyms. OtherDetails describes any other information. The Reference subconcept lists the various literature citations of the protein or protein complex described by the instances of CitationTitle, CitationAuthors, CitationEditors, CitationPublication, CitationReference, and CitationReferenceNumbers. A typical instance of Entry is as follows:

```
<Entry>
<ProteinOntologyID>PO0000000007</ProteinOntologyID>
<EntrySuperFamily>HUMAN</EntrySuperFamily>
<EntryFamily>PRION PROTEINS</EntryFamily>
<SourceDatabaseID>1E1P</SourceDatabaseID>
<SourceDatabaseName>PROTEIN DATA BANK</SourceDatabaseName>
<SubmissionDate>09-MAY-00</SubmissionDate>
<Title>HUMAN PRION PROTEIN VARIANT S170N</Title>
<Authors>L.CALZOLAI, D.A.LYSEK, P.GUNTERT, C.VON SCHROETTER,
R.ZAHN, R.RIEK, K.WUTHRICH</Authors>
<Experiment>NMR, 20 STRUCTURES</Experiment>
<Keywords>PRION PROTEIN</Keywords>
<CitationTitle>NMR STRUCTURES OF THREE SINGLE-RESIDUE VARIANTS OF
THE HUMAN PRION PROTEIN</CitationTitle>
<CitationAuthors>L.CALZOLAI, D.A.LYSEK, P.GUNTERT, C.VON
SCHROETTER, R.ZAHN, R.RIEK, K.WUTHRICH</CitationAuthors>
<CitationPublication>PROC.NAT.ACAD.SCI.USA</CitationPublication>
<CitationReference>V. 97 8340 2000</CitationReference>
<CitationReferenceNumbers>ASTM PNASA6 US ISSN
0027-8424</CitationReferenceNumbers>
</Entry>
```

5.5.5 Structure Concept

Structure Concept describes the protein structure details. Structure has two subconcepts: ATOMSequence and UnitCell. ATOMSequence is an example of the reuse of concepts in PO; it is constructed using generic concepts of Chain, Residue, and Atom. 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. Structure Concept defines ATOMSequence, with references to definitions of Chain and Residues, as:

```
<ATOMSequence>
<ProteinOntologyID>PO0000000004</ProteinOntologyID>
<Chain>
<AtomChain>A</AtomChain>
<Residue>
<ATOMResidue>ARG</ATOMResidue>
<Atom>
<AtomID>364</AtomID>
<Symbol>HE</Symbol>
<ATOMResSeqNum>148</ATOMResSeqNum>
-23.549</X>
<Y>3.766</Y>
<Z>-0.325</Z>
<Occupancy>1.0</Occupancy>
<TempratureFactor>0.0</TempratureFactor>
<Element>H</Element>
</Atom>
</Residue>
</Chains>
</ATOMSequence>
```

Protein crystallography data like a, b, c, alpha, beta, gamma, z, and SpaceGroup is described in *UnitCell* Concept.

5.5.6 StructuralDomains Concept

Structural folds and domains defining secondary structures of proteins are defined in StructuralDomains Concept. The subconcepts SuperFamily and Family of generic concept Family are used for identifying the protein family here. The subconcepts of StructuralDomains are Helices, Sheets, and OtherFolds. Helix, which is a subconcept of Helices, identifies a helix using HelixNumber, HelixID, HelixClass, and HelixLength Instances. Helix has a subconcept HelixStructure which gives the detailed composition of the helix. A typical instance of Helices Concept is:

```
<Helices>
<ProteinOntologyID>PO0000000002</ProteinOntologyID>
<StructuralDomainSuperFamily>HAMSTER</StructuralDomainSuperFamily>
<StructuralDomainFamily>PRION PROTEINS</StructuralDomainFamily>
<Helix>
<HelixID>1</HelixID>
<HelixNumber>1</HelixNumber>
<HelixClass>Right Handed Alpha</HelixClass>
<HelixLength>10</HelixLength>
<HelixStructure>
<HelixChain>A</HelixChain>
<HelixInitialResidue>ASP</HelixInitialResidue>
<HelixInitialResidueSeqNum>144</HelixInitialResidueSeqNum>
<HelixEndResidue>ASN</HelixEndResidue>
<HelixEndResidueSeqNum>153</HelixEndResidueSeqNum>
</HelixStructure>
</Helix>
</Helices>
```

Other secondary structures like sheets and turns (or loops) are represented using concepts of chains and residues in a similar way. Sheets has a subconcept Sheet which describes a sheet using SheetID and NumberStrands. Sheet has a subconcept Strands which describes the detailed structure of a sheet. A typical instance of Sheets Class is:

```
<Sheets>
<ProteinOntologyID>PO0000000001</ProteinOntologyID>
<StructuralDomainSuperFamily>MOUSE</StructuralDomainSuperFamily>
<StructuralDomainFamily>PRION PROTEINS</StructuralDomainFamily>
<Sheet>
<SheetID>S1</SheetID>
<NumberStrands>2</NumberStrands>
<Strands>
<StrandNumber>2</StrandNumber>
<StrandChain>NULL</StrandChain>
<StrandIntialResidue>VAL</StrandIntialResidue>
<StrandIntialResidueSeqNum>161</StrandIntialResidueSeqNum>
<StrandEndResidue>ARG</StrandEndResidue>
<StrandEndResidueSeqNum>164</StrandEndResidueSeqNum>
<StrandSense>ANTI-PARALLEL</StrandSense>
</Strands>
</Sheet>
</Sheets>
```

5.5.7 FunctionalDomains Concept

PO has the first Functional Domain Classification Model defined using FunctionalDomains Concept using: (1) data about cellular and organism source in SourceCell subconcept, (2) data about biological functions of protein in BiologicalFunction subconcept, and (3) data about active binding sites in proteins in ActiveBindingSites subconcept. Like StructuralDomains Concept, SuperFamily and Family subconcepts of generic concept Family are used for identifying the protein family here. SourceCell specifies a biological or chemical source of each biological molecule (defined by Molecule Concept earlier) in the protein. Biological functions of the protein complex are described in BiologicalFunction. BiologicalFunction has two subconcepts, PhysiologicalFunctions and PathologicalFunctions, and each of these has several subconcepts and sub-subconcepts describing various corresponding functions. The third subconcept of FunctionalDomains is ActiveBindingSites which has details about active binding sites in the protein. Active binding sites are represented in our ontology as a collection of various site groups, defined in SiteGroup Concept. SiteGroup has details about each of the residues and chains that form the binding site. There can be a maximum of seven site groups defined for a protein complex in PO. A typical instance of SourceCell in FunctionalDomains is:

```
<SourceCell>
<ProteinOntologyID>PO0000000009</ProteinOntologyID>
<SourceMoleculeID>1</SourceMoleculeID>
<OrganismScientific>HOMO SAPIENS</OrganismScientific>
<OrganismCommon>HUMAN</OrganismCommon>
<ExpressionSystem>ESCHERICHIA COLI; BACTERIA</ExpressionSystem>
<ExpressionSystemVector>PLAMID</ExpressionSystemVector>
<Plasmid>PRSETB</Plasmid>
</SourceCell>
```

5.5.8 ChemicalBonds Concept

Various chemical bonds used to bind various substructures in a complex protein structure are defined in ChemicalBonds Concept. Chemical bonds that are defined in PO by their respective subconcepts are: DisulphideBond, CISPeptide, HydrogenBond, ResidueLink, and SaltBridge. These are defined using generic concepts of Bind and AtomicBind. The chemical bonds that have binding residues (DisulphideBond, CISPeptide) reuse the generic concept of Bind. In defining the generic concept of Bind in protein ontology we again reuse the generic concepts of Chains and Residues. Similarly, the chemical bonds that have binding atoms (HydrogenBond, ResidueLink, and SaltBridge) reuse the generic concept of AtomicBind. In defining the generic concept of AtomicBind we reuse the generic concepts of Chains, Residues, and Atoms. A typical instance of a ChemicalBond is:

```
<CISPeptides>
<ProteinOntologyID>PO0000000003</ProteinOntologyID>
<BindChain1>H</BindChain1>
<BindResidue1>GLU</BindResidue1>
<BindResSeqNum1>145</BindResSeqNum1>
<BindChain2>H</BindChain2>
<BindResidue2>PRO</BindResidue2>
<BindResSeqNum2>146</BindResSeqNum2>
<AngleMeasure>-6.61</AngleMeasure>
<Model>0</Model>
</CISPeptides>
```

5.5.9 Constraints Concept

Various constraints that affect final protein conformation are defined in Constraints Concept using ConstraintID and ConstraintDescription. The constraints currently described in PO are as follows: (1) monogenetic and polygenetic defects present in genes that are present in molecules making proteins in the GeneDefects subconcept, (2) hydrophobicity properties in the Hydrophobicity concept, and (3) modification in residue sequences due to chemical environment and mutations in the ModifiedResidue concept. Posttranslational residue modifications are comprised of those amino acids that are chemically changed in such way that they could not be restored by physiological processes, as well as other rare amino acids that are translationally incorporated but for historical reasons are represented as modified residues. The RESID Database [41] is the most comprehensive collection of annotations and structures for protein modifications. The current version of RESID maps posttranslational modifications to both PIR and Swiss-Prot. Data in the GeneDefects class is entered as instances of GeneDefects class and is normally taken from OMIM [37] or scientific literature. A typical instance of a Constraint is:

```
<Constraints>
<ProteinOntologyID>PO0000000001</ProteinOntologyID>
<ConstraintID> 3 </ConstraintID>
<ConstraintDescription> MODIFICATION OF RESIDUES DUE TO
GLYCOSYLATION</ConstraintDescription>
</Constraints>
```

The complete class hierarchy of PO is shown in Figure 5.1. More details about PO are available at the Web site: http://www.proteinontology.info/.

5.5.10 Comparison with Protein Annotation Frameworks

In this section we compare the frameworks of our PO with PRONTO and PDBML.

5.5.10.1 PRONTO and PO

Machine-generated protein ontology generated by PRONTO is just a set of terms and relationships between those terms. PRONTO-generated ontology does not cover and map all the stages of the proteomics process from protein's primary structure to protein's quaternary structure. PRONTO uses iProLink literature-mining ontology to search and identify protein names in the MEDLINE database of biological literature. It then cross-references EBI's UniProt database to define relationships between these terms. PO, on the other hand, integrates data representation frameworks of various protein data sources—PDB, SCOP, RESID, and OMIM—to provide a unified vocabulary covering all the stages of proteomics process. PRONTO represents only two relationships between the terms of the ontology: is-a relation and part-of relation. Whereas PO represents five different relationships between the terms used in the ontology definition. They are: *SubConceptOf*, *PartOf*, *AttributeOf*, *InstanceOf*, and *ValueOf*.

5.5.10.2 PDBML and PO

PDBML is a XML Schema mapping the PDB Exchange Dictionary. In 2004, we did quite similar work [15–17] to PDBML by creating a XML Schema and RDF Schema mapping of PDB, Swiss-Prot, and PIR databases. PDBML lacks the hierarchical relationships as it is linked to the logical representation of PDB. The semantics of

ProtonOntology

- 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 5.1 Concept hierarchy of protein ontology.

data is preserved and translation from PDB to XML Schema is simple, but it cannot be used to process the content. PO with the power of OWL has no limitations in processing the content.

5.6 Protein Ontology Instance Store

The Protein Ontology Instance Store is created for entering existing protein data using the PO format. PO provides a technical and scientific infrastructure to allow evidence-based description and analysis of relationships between proteins. PO uses data sources including 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 various published scientific literature in various journals. The PO Instance Store is represented using OWL. The PO Instance Store at the moment contains data instances of following protein families: (1) Prion Proteins, (2) B.Subtilis, and (3) Chlorides. More protein data instances will be added as PO becomes more developed. All the PO instances are available for download (http://proteinontology.info/proteins.htm) in OWL format which can be read by any popular editor like Protégé (http://protege.stanford.edu/).

5.7 Strengths and Limitations of Protein Ontology

PO provides a unified vocabulary for capturing declarative knowledge about protein domain and for classifying that knowledge to allow reasoning. Information captured by PO is classified in a rich hierarchy of concepts and their interrelationships. 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 from defined generic concepts. The concepts derived from generic concepts are placed precisely into the concept hierarchy of PO to completely represent information that defines a protein complex.

As the OWL representation used in PO is an XML-Abbrev based (i.e., Abbreviated XML Notation), it can be easily transformed to the corresponding RDF and XML formats without much effort using the available converters. The PO Instance Store currently 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 will provide a specific set of rules to cover these application-specific semantics over the PO framework. The rules use only the relationships whose semantics are predefined to establish correspondence among terms in PO. These rules will help in defining semantic query algebra for PO to efficiently reason and query the underlying instance store.

For protein functional classification, in addition to the presence of domains, motifs or functional residues, the following factors are relevant: (1) similarity of three-dimensional protein structures, (2) proximity to genes (this may indicate that the proteins they produce are involved in same pathway), (3) metabolic functions of organisms, and (4) evolutionary history of the protein. At the moment, PO's functional domain classification does not address the issues of proximity of genes and evolutionary history of proteins. These factors will be added in the future to complete the functional domain classification system in PO. Also, the constraints defined in PO are not mapped back to the protein sequence, structure, and function they affect. Achieving this in the future will interlink all the concepts of PO.

The limitations of PO in terms of defining new concepts for protein functions and constraints on protein structure do not limit the use of generalized concepts in PO to define any kind of complex concept for proteomics research in the future.

5.8 Summary

Our protein ontology 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 generic concepts defined in PO which provide notions of classification, reasoning, and consistency when defining new concepts.

References

[1] Apweiler, R., et al., "UniProt: The Universal Protein Knowledgebase," *Nucleic Acids Res.*, Vol. 32, 2004, pp. 115–119.

[2] Apweiler, R., et al., "Protein Sequence Annotation in the Genome Era: The Annotation Concept of SWISS-PROT + TREMBL," *5th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Halkidiki, 1997.

[3] Boeckmann, B., et al., "The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 365–370.

[4] Berman, H., et al., "The Protein Data Bank and the Challenge of Structural Genomics," *Nature Structural Biology*, Structural Genomics Supplement, 2000, pp. 957–959.

[5] Bhat, T. N., et al., "The PDB Data Uniformity Project," *Nucleic Acids Res.*, Vol. 29, 2001, pp. 214–218.

[6] Weissig, H., and P. E. Bourne, "Protein Structure Resources," *Biological Crystallography*, Vol. D58, 2002, pp. 908–915.

[7] Westbrook, J., et al., "The Protein Data Bank: Unifying the Archive," *Nucleic Acids Res.*, Vol. 30, 2002, pp. 245–248.

[8] Mitra, P., and G. Wiederhold, "An Algebra for Semantic Interoperatibility of Information Sources," *Infolab*, Stanford University, Stanford, CA, 2001.

[9] Sidhu, A. S., T. S. Dillon, and E. Chang, "Ontological Foundation for Protein Data Models," *1st 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, 2005.

[10] Sidhu, A. S., T. S. Dillon, and E. Chang, "An Ontology for Protein Data Models," *27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2005 (IEEE EMBC 2005)*, Shanghai, China, 2005.

[11] Sidhu, A. S., T. S. Dillon, and E. Chang, "Advances in Protein Ontology Project," *19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006)*, Salt Lake City, UT, 2006.

[12] Sidhu, A. S., T. S. Dillon, and E. Chang, "Integration of Protein Data Sources Through PO," *17th International Conference on Database and Expert Systems Applications (DEXA 2006)*, Poland, 2006.

[13] Sidhu, A. S., T. S. Dillon, and E. Chang, "Towards Semantic Interoperability of Protein Data Sources," *2nd IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2006)* in conjunction with OTM 2006, France, 2006.

[14] Sidhu, A. S., et al., "Protein Ontology: Vocabulary for Protein Data," *3rd International IEEE Conference on Information Technology and Applications, 2005 (IEEE ICITA 2005)*, Sydney, 2005.

[15] Sidhu, A. S., et al., "Comprehensive Protein Database Representation," *8th International Conference on Research in Computational Molecular Biology 2004 (RECOMB 2004)*, San Diego, CA, 2004.

[16] Sidhu, A. S., et al., "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.

[17] Sidhu, A. S., et al., "An XML Based Semantic Protein Map," *5th International Conference on Data Mining, Text Mining and their Business Applications (Data Mining 2004)*, Malaga, Spain, 2004.

[18] Ashburner, M., et al., "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, Vol. 11, 2001, pp. 1425–1433.

[19] Lewis, S. E., "Gene Ontology: Looking Backwards and Forwards," *Genome Biology*, Vol. 6, 2004, pp. 103.1–103.4.

[20] Altmann, R. B., et al., "RiboWeb: An Ontology-Based System for Collaborative Molecular Biology," *IEEE Intelligent Systems*, 1999, pp. 68–76.

[21] Westbrook, J., et al., "PDBML: The Representation of Archival Macromolecular Structure Data in XML," *Bioinformatics*, Vol. 21, 2005, pp. 988–992.

[22] Mani, I., et al., "PRONTO: A Large-Scale Machine-Induced Protein Ontology," *2nd Standards and Ontologies for Functional Genomics Conference (SOFG 2004)*, Philidelphia, PA, 2004.

[23] Berman, H., K. Henrick, and H. Nakamura, "Announcing the Worldwide Protein Data Bank," *Nature Structural Biology*, Vol. 12, 2003, p. 980.

[24] Wu, C. H., et al., "The Protein Information Resource," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 345–347.

[25] Sasson, O., et al., "ProtoNet: Hierarchical Classification of the Protein Space," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 348–352.

[26] Bateman, A., et al., "The Pfam Protein Families Database," *Nucleic Acids Res.*, Vol. 30, 2002, pp. 276–280.

[27] Corpet, F., et al., "ProDom and ProDom-CG: Tools for Protein Domain Analysis and Whole Genome Comparisons," *Nucleic Acids Res.*, Vol. 28, 2000, pp. 267–269.

[28] Falquet, L., et al., "The Prosite Database, Its Status in 2002," *Nucleic Acids Res.*, Vol. 30, 2002, pp. 235–238.

[29] Attwood, T. K., et al., "PRINTS and Its Automatic Supplement, prePRINTS," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 400–402.

[30] Lo Conte, L., et al., "SCOP Database in 2002: Refinements Accommodate Structural Genomics," *Nucleic Acids Res.*, Vol. 30, 2002, pp. 264–267.

[31] Pearl, F. M. G., et al., "The CATH Database: An Extended Protein Family Resource for Structural and Functional Genomics," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 452–455.

[32] Huang, H., et al., "iProClass: An Integrated Database of Protein Family, Function and Structure Information," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 390–392.

[33] Mulder, N. J., et al., "The InterPro Database, 2003 Brings Increased Coverage and New Features," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 315–318.

[34] Letunic, I., et al., "Recent Advancements to the SMART Domain-Based Sequence Annotation Resource," *Nucleic Acids Res.*, Vol. 30, 2002, pp. 242–244.

[35] Haft, D., et al., "TIGRFAMs: A Protein Family Resource for Functional Identification of Proteins," *Nucleic Acids Res.*, Vol. 29, 2001, pp. 41–43.

[36] Dayhoff, M. O., "The Origin and Evolution of Protein Superfamilies," *Fed. Proc.*, Vol. 35, 1976, pp. 2132–2138.

[37] McKusick, V. A., *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*, 12th ed., Baltimore, MD: Johns Hopkins University Press, 1998.

[38] Frazier, M. E., et al., "Realizing the Potential of Genome Revolution: The Genomes to Life Program," *Science*, Vol. 300, 2003, pp. 290–293.

[39] Frazier, M. E., et al., "Setting Up the Pace of Discovery: The Genomes to Life Program," *2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)*, Stanford, CA, 2003.

[40] Collins, F. S., M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology," *Science*, Vol. 300, 2003, pp. 286–290.

[41] Garavelli, J. S., "The RESID Database of Protein Modifications: 2003 Developments," *Nucleic Acids Res.*, Vol. 31, 2003, pp. 499–501.