2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)
© 2008 IEEE.

# Current Status of Biomedical Ontologies: Developments in 2007

Amandeep S. Sidhu, Member, IEEE, Tharam S. Dillon, Fellow, IEEE and Elizabeth Chang, Member, IEEE

Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology Perth
e-mail: (Amandeep.Sidhu, Tharam.Dillon, Elizabeth.Chang)@cbs.curtin.edu.au

*Abstract*—**The goal of this paper is to survey existing biomedical ontologies and their developments in 2007. This paper discusses features of biomedical ontologies that allow true information integration in biomedical domain. The paper is compilation of several biomedical ontologies like Gene Ontology, Protein Ontology, etc. that have developed serving primarily the purposes of information extraction from on-line biomedical literature and databases.**

*Index Terms*— **Biomedical Ontologies, Biomedical Systems, Bioinformatics**

## I. Introduction

Biologists, in attempting to answer a specific biological question, now frequently choose their direction and select their experimental strategies by way of an initial computational analysis. Naturally, computers and computer tools are used to collect and analyse the results of largely automated instruments used in biological sciences. However, far more pervasive than this type of a requirement, the very nature of the intellectual discovery process requires access to the latest version of the worldwide collection of data, and the fundamental tools of bioinformatics that now are increasingly a part of the experimental methods themselves. A driving force in the life science discovery process is turning complex, heterogeneous data into useful, organized information and ultimately into systematized knowledge. The endeavour is simply the classic pathway for all science, Data ⇒ Information ⇒ Knowledge ⇒ Discovery, which earlier in the history of biology required only brainpower and pencil and paper, now requires sophisticated computational technology.

The problem of management of biological macromolecular data is as old as the data themselves. In 1998, a special issue of Nucleic Acids Research listed 64 different databanks covering diverse areas of biological research, and the nucleotide sequence data alone at over 1 billion bases. It is not only the flood of information and heterogeneity that make the issues of information representation, storage, structure, retrieval and interpretation critical. There also has been a change in the community of users. In the middle 1980s, fetching a biological entry on a mainframe computer was an adventurous step that only few dared. Now, at the end of the 1990s, thousands of researchers make use of biological databanks on a daily basis to answer queries, e.g. to find sequences similar to a newly sequenced gene, or to retrieve bibliographic references, or to investigate fundamental problems of modern biology [1]. New technologies, of which the World Wide Web (WWW) has been the most revolutionary in terms of impact on science, have made it possible to create a high density of links between databanks. Database systems today are facing the task of serving ever increasing amounts of data of ever growing complexity to a user community that is growing nearly as fast as the data, and is becoming increasingly demanding.

## II. Ontology Preliminaries

In response to current advances in technology and the increasing scope of research, massive amounts of data are routinely deposited in public and private databases. The scope of public data sources ranges from the comprehensive, multidisciplinary, community informatics centres, supported by public funds and sustained by team of specialists, to small boutique data sources provided by individual investigators. The content of the databases varies greatly, reflecting the broad disciplines and sub-disciplines across life sciences from molecular biology and cell biology, to medical and clinical trials, to ecology and biodiversity. In this section, we briefly discuss various features of biological databases and then give samples of various public biological databases.

Biological data sources represent a loose collection of autonomous web sites, each with its own governing body and infrastructure. These sites vary in almost every possible instance such as computing platform, access and data management system. Much of the available biological data exists in legacy systems in which there are no structured information management systems. These data sources are inconsistent at the semantic level, and more often than not, there is no adequate meta-data specification. Until recently, biological databases were not designed for interoperability [2]. Data elements in public or proprietary databases are stored in heterogeneous formats ranging from simple files to fully structured database systems that are often ad hoc, application-specific or vendor-specific. For example, scientific literature, images, and other free-text documents are commonly stored in unstructured or semi-structured formats (plain text, HTML, XML). Genomic, microarray gene expression, and proteomic data are stored in conventional flat files and spreadsheet programs or in structured relational databases (Oracle, Sybase, DB2, and Informix).

Perhaps the technical problems of standardization discussed in the preceding paragraphs could be addressed more easily in the context of a more general logical structure. As noted by Hafner [3], general biological data resources are databases rather than knowledge bases: they describe miscellaneous objects according to the database schema, but no representation of general concepts and their relationships is given. Schulze-Kremer [4] addressed this problem by developing ontologies for knowledge sharing in molecular biology. He proposed to create a repository of terms and concepts relevant to molecular biology, hierarchically organized by means of 'is a subset of' and 'is member of' operators.

## III. Biomedical Ontologies

The term ontology is originally a philosophical term referred as *"the object of existence"*. Computer Science community borrowed the term ontology to refer to a "specification of conceptualisation" for knowledge sharing in artificial intelligence [5]. Ontologies provide a conceptual framework for a structured representation of the meaning, through a common vocabulary, on a given domain — in this case, biological or medical— that can be used by either humans or automated software agents on a the domain. This shared vocabulary usually includes concepts, relationships between concepts, definitions for these concepts and relationships and also the possibility of defining ontology rules and axioms; in order to define a mechanism to control the objects that can be introduced in the ontology and to apply logical inference. Ontologies in biomedicine have emerged because of the need for common language for effective communication across diverse sources of biological data and knowledge.

Several Biomedical Ontologies like UMLS [6] Gene Ontology [7], Protein Ontology [8], MGED Ontology [9], and TAMBIS Ontology [10] have developed, often reflecting mere relations of 'association' between what are called 'concepts', and serving primarily the purposes of information extraction from on-line biomedical literature and databases. In recent years, we have learned a great deal about the criteria, which must be satisfied if ontology is to allow true information integration and automatic reasoning across data and information derived from different sources. Substantial contributions have been carried out in medicine for the development of standards, medical terminologies and coding systems. The most important one, from the ontological perspective, is the MeSH (Medical Subject Headings) ontology, used to index Medline documents. MeSH [11] by the National Library of Medicine (NLM) mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several different types of terms, such as Descriptor, Qualifiers, Publication Types, Geographics, and Entry terms. MeSH has got more than 18000 categories, with a poly tree based, hierarchical structure where a term can appear in different branches. In 1986, NLM began a long-term goal to build Unified Medical Language System (UMLS). UMLS [6, 12, 13] is a repository of biomedical vocabularies and is NLM's biomedical ontology. The purpose of the UMLS is to improve the ability of computer programs to understand biomedical meaning and to use its understanding to retrieve relevant machine readable information for users [13]. The UMLS integrates over 2 million names for some 900,000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts.

Gene Ontology [7, 14] consortium lead ontological development in the genetic area. 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. The GO Consortium was initially a collaboration among Mouse Genome Database [15], FlyBase [16], and Saccharomyces Genome database [17] efforts. GO is now a part of UMLS, and the GO Consortium is a member of the Open Biological Ontologies consortium to be discussed later in this section. 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 other as well [18]. In this way, functional predictions can be made by applying prior knowledge to infer function of novel entity (either a gene or a protein). GO consists of three distinct ontologies, each of which serves as an organizing principle for describing gene products. The intention is that each gene product should be annotated by classifying it three times, once within each ontology [19].

GO is the result of the effort to enumerate and model concepts used to describe genes and gene products. The central unit for description in GO is a *concept*. Concept consists of unique identifier and one or more strings (referred to as *terms*) that provide a controlled vocabulary for unambiguous and consistent naming. Concepts exist in a hierarchy of IsA and PartOf relations in a directed acyclic graph (DAG) that locates all concepts in the knowledge model with respect to their relationships with other concepts. More details about Gene Ontology are at: http://www.geneontology.org/

We are building Protein Ontology [20-23] 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 consists of concepts (or classes), which are data descriptors for proteomics data and the relationships among these concepts. PO has (1) a hierarchical classification of concepts represented as classes, from general to specific; (2) a list of attributes related to each concept, for each class; (3) a set of relationships between classes to

2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)

link concepts in ontology in more complicated ways then implied by the hierarchy, to promote reuse of concepts in the ontology; and (4) a set of algebraic operators for querying protein ontology instances. More details about Protein Ontology are at: http://www.proteinontology.org.au/

The MGED Ontology (MO) is developed by 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 analyze the data [9]. 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; it is not an ontology of molecular, cellular, or organism biology, such as the Gene Ontology. MO version 1.2 contains 229 classes, 110 properties and 658 instances.

TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) uses an ontology to enable biologists to ask questions over multiple external databases using a common query interface [24]. The TAMBIS ontology (TaO) [10] describes a wide range of bioinformatics tasks and resources, and has a central role within the TAMBIS system. An interesting difference between the TaO and some of the other ontologies is that the TaO does not contain any instances. The TaO only contains knowledge about bioinformatics and molecular biology concepts and their relationships - the instances they represent still reside in the external databases. The TaO is a dynamic ontology, in that it can grow without the need for either conceptualizing or encoding new knowledge.

The National Center for Biomedical Ontology is an NIH National Center for Biomedical Computing (NCBC): a consortium comprised of leading biologists, clinicians, informaticians, and ontologists who are working together to develop innovative technology and methods that allow scientists to record, manage, and disseminate biomedical information and knowledge in machine-processable form.

The Center is developing two major repositories of biomedical content: (1) Open Biomedical Ontologies (OBO), a comprehensive, online library of open-content ontologies and controlled terminologies, and (2) Open Biomedical Data (OBD), a database resource that will allow expert scientists to archive experimental data that is fully described (annotated) using the OBO ontologies and terminologies. The biomedical research community will access OBO and OBD via a system called BioPortal. List of Biomedical Ontologies and more details about them available at National Center for Biomedical Ontologies is available at: http://cbioapprd.stanford.edu/ncbo/faces/pages/ontology_list.xhtml

## IV. Open Issues in Biomedical Ontologies

Research into different biological systems uses different organisms chosen specifically because they are amenable to advancing these investigations. For example, the rat is a good model for the study of human heart disease, and the fly is a good model for the study of cellular differentiation. For each of these model systems, there is a database employing curators who collect and store the body of biological knowledge for that organism. Mining of Scientific Text and Literature is done to generate a list of keywords that are used as GO terms. However, querying heterogeneous, independent databases in order to draw these inferences is difficult: The different database projects may use different terms to refer to the same concept and the same terms to refer to different concepts. Furthermore, these terms are typically not formally linked with each other in any way. GO seeks to reveal these underlying biological functionalities by providing a structured controlled vocabulary that can be used to describe gene products, and is shared between biological databases. This facilitates querying for gene products that share biologically meaningful attributes, whether from separate databases or within the same database.

Association between ontology nodes and proteins, namely, protein annotation through gene ontology, is an integral application of GO. To efficiently annotate proteins, the GO Consortium developed a software platform, the GO Engine, which combines rigorous sequence homology comparison with text information analysis. During evolution, many new genes arose through mutation, duplication, and recombination of the ancestral genes. When one species evolved into another, the majority of orthologs retained very high levels of homology. The high sequence similarity between orthologs forms one of the foundations of the GO Engine. Text information related to individual genes or proteins is immersed in the vast ocean of biomedical literature. Manual review of the literature to annotate proteins presents a daunting task. Several recent papers described the development of various methods for the automatic extraction of text information [25, 26]. However, the direct applications of these approaches in GO annotation have been minimal. A simple correlation of text information with specific biological ontology nodes in the training data should predict association for unannotated biomedical data. Correlation methodology should combine homology information, a unique data-clustering procedure, and text information analysis to create the best possible annotations.

For Protein Functional Classification, in addition to the presence of domains, motifs or functional residues, the following factors are relevant: (a) similarity of three dimensional protein structures; (b) proximity to genes (which may indicate that the proteins they produce are involved in the same pathway); (c) metabolic functions of organisms; and (d) 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 to complete the Functional Domain Classification System in a future version of the Protein Ontology. Also, more constraints need to be added to Protein Ontology other than the three existing types of constraints, by studying the effects of environment on various protein families that exist. Also, the constraints defined in PO are not mapped back to the protein sequence, structure and function that they affect. Achieving this in future will inter-link all the concepts of PO. We also need to analyse all the protein families to study how the Protein Ontology representation will help us to understand in detail the linkage between various protein families.

Currently, we are developing a trustworthy ontology [27] to automate the process of additions and modifications to the Protein Ontology through online interfaces. This also assists in providing a degree of separation between the entered concepts and the actual protein ontology available to the users through the use of an Intermediate Protein Ontology (IPO). A trustworthy Protein Ontology framework will ensure that only valid and correct concepts are added to Protein Ontology. A PO administrator uses an administration console to skim through IPO using a defined set of rules that denotes what a correct concept would be, what a correct relationship between different concepts would be, and what a correct instance of the concept would be [28]. These sets of rules utilize structure and semantics of the PO to facilitate validation of any changes made to the IPO by research assistants. To identify the pattern of correctness of the information that has been added, we calculate a correctness value [29] for every concept entered by the researcher into the Intermediate Protein Ontology (IPO). For a correct concept entered, 0.1 increases the reputation; whereas for an incorrect concept entered, the 0.05 decreases the reputation. The final value of reputation for each set of entries by a researcher determines the correctness of information entered by the researchers.

## V. References

[1] E. V. Koonin and M. Y. Galperin, "Prokaryotic genomes: the emerging paradigm of genome-based microbiology," *Current Opinons in Genetic Development,* vol. 7, pp. 757-763, 1997.

[2] P. Karp, "Database links are a foundation for interoperability," *Trends in Biotechnology,* vol. 14, pp. 273-279, 1996.

[3] C. D. Hafner and N. Fridman, "Ontological foundations for biology knowledge models," in *4th International Conference on Intelligent Systems for Molecular Biology*, St. Louis, 1996, pp. 78-87.

[4] S. Schulze-Kremer, "Ontologies for Molecular Biology," in *Pacific Symposium of Biocomputing*, Hawaii, 1998, pp. 693-704.

[5] T. R. Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing," *International Journal of Human and Computer Studies,* vol. 43, pp. 907-928, 1995.

[6] K. Baclawski, J. Cigna, M. M. Kokar, P. Magner, and B. Indurkhya, "Knowledge Representation and Indexing Using the Unified Medical Language System," in *Pacific Symposium on Biocomputing*, Honolulu, Hawaii, 2000, pp. 490-501.

[7] M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, J. C. Cherry, J. Corradi, and K. Dolinski, "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research,* vol. 11, pp. 1425-1433, 2001.

[8] A. S. Sidhu, T. S. Dillon, and E. Chang, "Ontological Foundation for Protein Data Models," in *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, pp. 916-925.

[9] P. L. Whetzel, H. Parkinson, H. C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S. Sansone, C. Taylor, J. White, and C. J. Stoeckert, "The MGED Ontology: a resource for semantics-based description of microarray experiments," *Bioinformatics,* vol. 22, pp. 866-873, 2006.

[10] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass, "An Ontology for Bioinformatics Applications," *Bioinformatics,* vol. 15, pp. 510-520, 1999.

[11] S. J. Nelson, D. Johnston, and B. L. Humphreys, "Relationships in Medical Subject Headings," in *Relationships in the organization of knowledge*, C. A. Bean and R. Green, Eds. New York: Kluwer Academic Publishers, 2001, pp. 171-184.

[12] C. Lindberg, "The Unified Medical Language System (UMLS) of the National Library of Medicine," *Journal of American Medical Record Association,* vol. 61, pp. 40-42, 1990.

[13] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The Unified Medical Language System," *Methods of information in medicine,* vol. 32, pp. 281-291, 1993.

[14] S. E. Lewis, "Gene Ontology: looking backwards and forwards," *Genome Biology,* vol. 6, pp. 103.1-103.4, 2004.

[15] J. A. Blake, J. T. Eppig, J. E. Richardson, and M. T. Davisson, "The Mouse Genome Database (MGD): a community resource. Status and enhancements. The Mouse Genome Informatics Group," *Nucleic Acids Research,* vol. 26, pp. 130-137, 1998.

[16] M. Ashburner, "FlyBase," *Genome News,* vol. 13, pp. 19-20, 1993.

[17] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. R. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, A. P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, and D. Nusbaum, "A gene map of the human genome," *Science,* vol. 274, pp. 540-

546, 1996.

[18] O. D. King, R. E. Foulger, S. Dwight, J. White, and F. P. Roth, "Predicting gene function from patterns of annotation," *Genome Research,* vol. 13, pp. 896-904, 2003.

[19] A. G. Fraser and E. M. Marcotte, "A probabilistic view of gene function," *Nature Genetics,* vol. 36, pp. 559-564, 2004.

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

[21] A. S. Sidhu, T. S. Dillon, and E. Chang, "Integration of Protein Data Sources through PO," in *17th International Conference on Database and Expert Systems Applications (DEXA 2006)*, Poland, 2006, pp. 519-527.

[22] A. S. Sidhu, T. S. Dillon, and E. Chang, "Protein Ontology," in *Biological Database Modeling (In Press)*, J. Chen and A. S. Sidhu, Eds. New York: Artech House, 2007, pp. 39-60.

[23] A. S. Sidhu, T. S. Dillon, B. S. Sidhu, and H. Setiawan, "An XML based semantic protein map," in *5th International Conference on Data Mining, Text Mining and their Business Applications (Data Mining 2004)*, Malaga, Spain, 2004, pp. 51-60.

[24] P. G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, and R. Stevens, "TAMBIS - transparent access to multiple bioinformatics information sources," in *6th International Conference on Intelligent Systems for Molecular Biology*, Montreal, Canada, 1998, pp. 25-34.

[25] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics,* vol. 28, pp. 21-28, 2001.

[26] Q. Li, P. Shilane, N. F. Noy, and M. A. Musen, "Ontology acquisition from on-line knowledge sources.," in *AMIA 2000 Annual Symposium*, Los Angeles, CA, 2000, pp. 497-501.

[27] F. K. Hussain, A. S. Sidhu, T. S. Dillon, and E. Chang, "Engineering Trustworthy Ontologies: Case Study of Protein Ontology," in *19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006)*, Salt Lake City, Utah, 2006, pp. 617-622.

[28] A. S. Sidhu, F. K. Hussain, T. S. Dillon, and E. Chang, "A Methodology for Protein Ontology Maintenance," in *The 33rd Annual Conference of the IEEE Industrial Electronics Society*, Taipei, Taiwan, 2007.

[29] A. S. Sidhu, F. K. Hussain, T. S. Dillon, and E. Chang, "Trust based Decision Making Approach for Protein Ontology," in *12th IEEE Conference on Emerging Technologies and Factory Automation*, Patras, Greece, 2007.