Current Status of Biomedical Ontologies: Developments in 2006

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 2006. This paper discusses features of biomedical ontologies that allow true information integration in biomedical domain. The paper is compilation of several biomedical ontologies like UMLS, Gene Ontology, Protein Ontology, MGED Ontology, and TAMBIS Ontology that 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.**

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

I. Introduction

Bioinformatics tools and systems perform a diverse range of functions including: data collection, data mining, data analysis, data management, and data integration. Computer-aided technology directly supporting medical applications is excluded from this definition and is refereed as medical informatics. This thesis is not an attempt at authoritatively describing the gamut of information contained in this field. Instead, it focuses on area of proteomics data integration, access, and interoperability as these areas form the cornerstone of the field. However, most of the approaches presented are generic integration systems that can be used in many similar contexts.

Since the first efforts of Maxam [1] and Sanger [2], the DNA sequence databases have been doubling in size every 18 months or so. This trend continues unabated. The problem of management of biological macromolecular sequence data is as old as the data themselves. In 1998, special issue of *Nucleic Acids Research* lists 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 and heterogeneity that make the issues of information representation, storage, structure, retrieval and interpretation critical. There also has been a change in user community. 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 [3]. 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 data, and is getting more and more demanding.

II. Need For Ontologies

Public databases distribute their contents as flat files, in some cases including indices for rapid data retrieval. In principle, all flat file formats are based on the organizational hierarchy of database, entry, and record. Entries are the fundamental entities of molecular databases, but in contrast to the situation in the living cell that they purport to describe, database entries store objects in the form of atomic, isolated, non-hierarchical structures. Different databases may describe different aspects of the same biological unit, e.g. the nucleic acid and amino acid sequences of a gene, and the relationship between them must be established by links that are not intrinsically part of the data archives themselves.

The development of individual databases has generated a large variety of formats in their implementations. There is consensus that a common language, or at least that mutual intelligibility, would be a good thing, but this goal has proved difficult to achieve. Attempts to unify data formats have included application of Backus–Naur based syntax [4], the development of an object-oriented database definition language [5] and the use of Abstract Syntax Notation 1 [6, 7]. None of these approaches has achieved the hoped for degree of acceptance. Underlying the questions of mechanisms of intercommunication between databases of different structure and format is the need for common semantic standards and controlled vocabulary in annotations [8, 9]. This problem is especially acute in comparative genomics. From the technological point of view, inter-genome comparisons are inter-database comparisons, which means that the databases to be compared have to speak the same language: keywords, information fields, weight factors, object catalogues, etc.

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 [10], 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 [11] addressed this problem by de-

1-4244-0470-3/07/$20.00 ©2007 IEEE

veloping 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

Existing traditional approaches do not address the complex issues of biological data discussed in earlier sections. However, recent work on ontologies intends to provide solutions to these issues. 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 [12]. 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 [13] Gene Ontology [14], Protein Ontology [15], MGED Ontology [16], and TAMBIS Ontology [17] have developed, often reflecting mere relations of 'association' between what are called 'concepts', and serving primarily the purposes of information extraction from online 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 [18] 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 [13, 19, 20] 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 [20]. 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. Vocabularies integrated in UMLS include the NCBI taxonomy, Gene Ontology (GO), the Medical Subject Headings (MeSH), OMIM and Digital Anatomist Symbolic Knowledge Base. UMLS concepts are not only interrelated, but may also be linked to external resources such as GenBank [21]. UMLS is composed of: the Metathesaurus (META), the SPECIALIST lexicon and associated lexical programs, and the Semantic Network (SN) [22].

In 1998, efforts to develop the Gene Ontology [14, 23] began, leading 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 [24], FlyBase [25], and Saccharomyces Genome database [26] 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 [27]. 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 [28]. The three GO ontologies are:

1. **Molecular Function:** This ontology describes the biochemical activity of gene product. For example, a gene product could be a transcription factor or DNA helicase. This classifies what kind of molecule the gene product is.
2. **Biological Process:** This ontology describes the biological goal to which a gene product contributes. For example, mitosis or purine metabolism. An ordered assembly of molecular functions accomplishes such a process. This describes what a molecule does or is involved in doing.
3. **Cellular Component:** This ontology describes the location in a cell in which the biological activity of the gene product is performed. Examples include the nucleus, telomere, or an origin recognition complex. This is where gene product is located.

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.

Eight years have now passed and GO has grown enormously. GO is now clearly defined and a model for numerous other biological ontology projects that aim similarly to achieve structured, standardized vocabularies for describing biological systems. GO is a structured network consisting of defined terms and the relationships between them that describe attributes of gene products. There are many measures demonstrating its success. Characteristics of GO that we believe are most responsible for its success: community involvement; clear goals; limited scope; simple, intuitive structure; continuous evolution; active curation; and early use. At present there are close to 300 articles in PubMed referencing GO. Among large institutional databanks, Swiss-Prot now uses GO for annotating the peptide sequences it maintains. The number of organism groups participating in the GO consortium has grown every quarter-year from the initial three to roughly two dozen. Every conference has talks and posters either referencing or utilizing GO, and within the genome community it has become the accepted standard for functional annotation. More details about Gene Ontology are at: http://www.geneontology.org/

We are building Protein Ontology [29-32] 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 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.info/

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 [16]. 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 [33]. The TAMBIS ontology (TaO) [17] 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.

IV. National Center for Biomedical Ontology

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 goals of the Center are: (1) to help unify the divergent and isolated efforts in ontology development by promoting open-source, standards-based tools to create, manage, and use ontologies, (2) to create new software tools to help scientists to use ontologies to annotate and analyze biomedical data, and (3) to provide a national resource for the ongoing evaluation, integration, and evolution of biomedical ontologies and associated tools and theories in the context of driving biomedical projects (DBPs). The National Center for Biomedical Ontology seeks to provide tools and methods to enhance the use of ontologies throughout biomedicine, and welcomes all kinds of collaborative projects that will benefit from the Center's resources and that will provide strong "applications pull" to stimulate the Center's ongoing research and development activities.

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—a Web site and a suite of Web services that will enable both human users and computer-based agents to access the rich content that the Center and its collaborators will curate. 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

V. Open Issues in Biomedical Ontologies

Research into different biological systems uses different organisms that are chosen 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 to study 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 list of keywords that is used as biomedical ontology 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. Biomedical Ontologies seek to reveal these underlying biological functionalities by providing a structured controlled vocabulary that can be used to describe gene products, and shared between biological databases. This facilitates querying for gene products that share biologically meaningful attributes, whether from separate databases or within the same database.

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 [34, 35]. 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.

VI. References

[1] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of National Academic of Science*, vol. 74, pp. 560-564, 1977.

[2] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *Journal of Molecular Biology*, vol. 94, pp. 441-448, 1975.

[3] 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.

[4] D. G. George, H.-W. Mewes, and H. Kihara, "A standardized format for sequence data exchange," *Protein Seq. Data Anal.*, vol. 1, pp. 27-29, 1987.

[5] D. G. George, B. C. Orcutt, H.-W. Mewes, and A. Tsugita, "An object-oriented sequence database definition language (sddl)," *Protein Seq. Data Anal.*, vol. 5, pp. 357-399, 1993.

[6] H. Ohkawa, J. Ostell, and S. Bryant, "MMDB: an ASN.1 specification for macromolecular structure," presented at 3rd International Conference on Intelligent Systems for Molecular Biology, Cambridge, United Kingdom, 1995.

[7] J. Ostell, "GenInfo ASN.1 Syntax: Sequences," National Library of Medicine, NIH 1, 1990.

[8] S. Pongor, "Novel databases for molecular biology," *Nature*, vol. 332, pp. 24-24, 1998.

[9] C. J. Rawlings, "Designing databases for molecular biology," *Nature*, vol. 334, pp. 447-447, 1998.

[10] C. D. Hafner and N. Fridman, "Ontological foundations for biology knowledge models," presented at 4th International Conference on Intelligent Systems for Molecular Biology, St. Louis, 1996.

[11] S. Schulze-Kremer, "Ontologies for Molecular Biology," presented at Pacific Symposium of Biocomputing, Hawaii, 1998.

[12] 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.

[13] K. Baclawski, J. Cigna, M. M. Kokar, P. Magner, and B. Indurkhya, "Knowledge Representation and Indexing Using the Unified Medical Language System," presented at Pacific Symposium on Biocomputing, Honolulu, Hawaii, 2000.

[14] 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.

[15] A. S. Sidhu, T. S. Dillon, and E. Chang, "Ontological Foundation for Protein Data Models," presented at 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminol-

ogy," *Nucleic Acids Research*, vol. 32, pp. 267-270, 2004.

[22] J. C. Denny, J. D. Smithers, and R. Miller, ""Understanding" medical school curriculum content using Knowledge Map," *Journal of the American Medical Informatics Association*, vol. 10, pp. 351-362, 2003.

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

[24] 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.

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

[26] 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.

[27] 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.

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

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

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

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

[32] A. S. Sidhu, T. S. Dillon, B. S. Sidhu, and H. Setiawan, "An XML based semantic protein map," presented at 5th International Conference on Data Mining, Text Mining and their Business Applications (Data Mining 2004), Malaga, Spain, 2004.

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

[34] 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.

[35] Q. Li, P. Shilane, N. F. Noy, and M. A. Musen, "Ontology acquisition from on-line knowledge sources.," presented at AMIA 2000 Annual Symposium, Los Angeles, CA, 2000.