Biomedical Ontologies: Session Introduction

O. Bodenreider, J.A.. Mitchell, and A.T. McCray

Pacific Symposium on Biocomputing 10:76-78(2005)

BIOMEDICAL ONTOLOGIES

O. BODENREIDER

U.S. National Library of Medicine
8600 Rockville Pike, MS 43, Bethesda, Maryland, 20894, USA
E-mail: olivier@nlm.nih.gov

J. A. MITCHELL

University of Missouri, Department of Health Management & Informatics,
Columbia, Missouri, 65211, USA
E-mail: MitchellJo@health.missouri.edu

A. T. MCCRAY

U.S. National Library of Medicine
8600 Rockville Pike, MS 52, Bethesda, Maryland, 20894, USA
E-mail: mccray@nlm.nih.gov

In the "post-genomic era", biomedical ontologies are becoming increasingly popular in the computational biology community as the focus of biology has started to shift from mapping genomes to analyzing the vast amount of information resulting from functional genomics research. In fact, biomedical ontologies play a central role in integrating the information about various model organisms, acquired under different conditions and stored in heterogeneous databases. The need for a controlled vocabulary to annotate gene products certainly explains the success of the Gene Ontology™ (GO), which has become a *de facto* standard in this domain.

The presence of research focused on or enabled by biomedical ontologies in molecular biology conferences illustrates the increasing role of ontologies in biological research. At the Pacific Symposium of Biocomputing (PSB), for example, the place of biomedical ontologies has grown from one paper in 1998 to 41 papers submitted to our session this year. Similarly, a large number of papers presented at the 12th conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2004) focused on some aspect of biomedical ontology. Finally, events such as the workshop on Bio-Ontologies collocated with ISMB each year since 1998 and the success of the Standards and Ontologies for Functional Genomics conferences are another testimony to the importance of ontologies to biologists.

While the purpose of biomedical *terminology* is to collect the names of entities (i.e., substances, qualities and processes) employed in the biomedical domain, the purpose of biomedical *ontology* is to study classes of entities in reality which are of biomedical significance. Beyond names, ontology is concerned with the principled

definition of biological classes and the relations among them. In practice, as they are more than lists of terms but do not necessarily meet the requirements of formal organization, the many products developed by biomedical terminologists and ontologists often constitute an "ontology gradient". Gene Ontology is one such structure lying between terminology and ontology.

Biomedical ontology research encompasses a variety of entities (from dictionaries of names for biological products, to controlled vocabularies, to principled knowledge structures) and processes (i.e., acquisition of ontological relations, integration of heterogeneous databases, use of ontologies for reasoning about biological knowledge). This session reflects many aspects of this research. Not surprisingly, a large number of submissions focus on the Gene Ontology.

A first group of papers investigates foundational issues in biomedical ontology as well as the creation of ontological resources. Hoffman et al. discuss the extension of existing clinical vocabularies to include molecular diagnostics and cytogenetics concepts, using information from the RefSeq database. Following-up on the study on the compositional structure of GO terms they presented last year at PSB, Ogren et al. reflect on the implications of such properties on the curation and usage of GO. In addition to lexical properties, Bodenreider et al. show that statistical methods applied to annotation databases can also help reveal associative relations among GO terms. Finally, Spasić et al. present a method for measuring similarity among biomedical terms, which not only utilizes ontological relations but can also contribute to identifying additional relations.

The second group of papers focuses on the role played by ontologies in integrating disparate biomedical resources. Marshall et al. explore five levels of constraint for matching biological entities and the links among them. Foreseeing what a biomedical Semantic Web would require, Bechhofer et al. developed a system which automatically adds semantic annotations to existing web resources, enabling the dynamic integration of such resources. In the tradition of the Microarray Gene Expression Data (MGED) ontology, Orchard et al. investigate the resources required for describing the complex experimental procedures used in proteomics and sharing the corresponding data. A practical application of integration is presented by Gennari et al., visualizing in the same information space anatomical data and various genomic resources.

The remaining papers have a somewhat different perspective on biomedical ontologies. While some of these papers make a limited use of the rich structure of ontologies and draw essentially on their terminology component, none of these papers could have existed without the standardization fostered by ontologies. Most papers, however, take advantage – to some degree – of the relations recorded in biomedical ontologies.

Two papers exploit the information contained in various annotation databases to investigate the relations among biological entities. Tari et al. study the properties of functionally-related gene networks. Xiong et al. use hyperclique patterns to identify

functional modules in protein complexes. Conversely, Yamakawa et al. analyze the common features in sets of genes using their annotations, through gene-GO term bipartite graphs.

Two papers focus on predicting the functional annotations of biological entities. Hayete and al. use decision trees to learn the associations between GO terms and protein domains. Lu and al. show that subcellular localization information can be predicted from molecular function information. Finally, with the GenesTrace system, Cantor et al. analyze the relations between diseases and genes as represented in existing databases integrated through terminology and ontology resources.

This session reflects the diversity of the biomedical ontology community. Topics of interest range from the foundational issues in defining the entities existing in biological reality to the formalisms required to represent these entities and their interrelations. Other topics involve the use of ontologies to enable sharing complex biological information and the integration of heterogeneous databases, as well as the various applications made possible by such integrated data repositories.

As better ontological resources are developed, such applications will increasingly enable complex reasoning about biomedical knowledge. As standard formalisms and communication protocols emerge, the use of heterogeneous resources will become more dynamic and automatic to biologists. Ultimately, the applications supported by biomedical ontologies will not only make it possible for biologists to keep up with an increasing amount of information, but hopefully also free them from the least interesting tasks. Beyond the *personal digital assistants* of today, which store our agendas and email messages, the *digital research assistants* of tomorrow will scan online information sources for us, summarize their content and organize the related knowledge into research hypotheses.