

## Principles of Ontology Construction

### *Motivation*

---

Biomedical advances are increasingly stymied, not by a lack of experimental data, but by a lack of data that is available in computable form. Biological data must be readily accessible, comparable, and fully correlated, and it must be captured in a language that allows the formulation of coherent testable hypotheses if it is to efficiently provide relevant answers to scientific inquiries and thus enable discoveries. The bottleneck in research today, in other words, is the task of sifting through large data sets to decide what experiments should be done.

Ontological frameworks can provide a shared language for communicating biological information and thereby integrating biological knowledge and removing this data bottleneck. These rigorous semantic descriptions of the entities and relationships between these entities can then be used to formulate hypotheses about and navigate through the volumes of data. To this end there are an increasing number of groups developing ontologies in assorted biological domains. However, these efforts will only be beneficial and aid biological data integration if certain criteria are met. These prerequisites are that the ontologies are non-overlapping, that they are accepted and used by the community, and that they are well-principled.

### *Goals and objectives*

---

This tutorial will provide instruction in all areas of ontology development, but the primary focus will be the fundamental principles and approaches that are required. First, we will provide background on existing ontological efforts and what is currently available for various biological domains. Second, we will provide instruction in the practical aspects of ontology development by describing available tools and their usage. Third and most importantly, we will offer training in the essential tried and tested principles that are required to build a robust formalization that corresponds well with reality. The goal of this tutorial is to foster communication and the adoption of best practices within the community, to encourage cooperative development, and to ensure that contributed biological ontologies are sufficiently well principled that they may be reasoned over within an open cumulatively growing framework.

### *Instructors*

---

The two instructors will be Suzanna Lewis and Barry Smith. The other individuals will assist in developing the curriculum and serve as teaching assistants during the tutorial (particularly important for answering questions during the interactive portions of the tutorial).

**Suzanna Lewis:** Head of the Berkeley *Drosophila* Genome Project bioinformatics group and a founder of the Gene Ontology Consortium. She, together with Martin Reese, taught one previous ISMB tutorial, the Genome Annotation Assessment Project held in 1999 in Heidelberg.

**Barry Smith:** Research Director of the European Centre for Ontological Research (ECOR), which aims at applying the theories developed by philosophers to a variety of problems in information science and related areas.

**Michael Ashburner:** former Joint Head of the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL). He also is Professor of Biology at the University of Cambridge. For the past 15 or so years, he has had a strong interest in the provision of infrastructure for databases for biologists. He is a founder of FlyBase, a major database for researchers using *Drosophila* as a model organism, and of the Gene Ontology Consortium.

**Mark Musen:** Head of Stanford Medical Informatics an academic program devoted to basic investigation and training in both clinical informatics and bioinformatics.

**Rama Balakrishnan:** Scientific Content Editor at the *Saccharomyces* Genome Database and for the Gene Ontology.

**David Hill:** Scientific Content Editor at the Mouse Genome Informatics and for the Gene Ontology.

### *Level*

---

Intermediate

### *Intended audience*

---

This tutorial is primarily aimed at bioinformaticists who are directly involved in developing or are interested in utilizing ontological techniques for managing biological data. This course will provide instruction in the available ontological methodologies as well as principles to use for developing content. This tutorial will also be of interest to those who are annotating data and need to know what semantic descriptions are available and what the implications of an annotation are. We will ask those who are developing ontologies to provide them beforehand for evaluation and case studies. It is expected that the audience have some existing familiarity with biological ontologies.

### *Length*

---

Half day.

### *Detailed outline of the presentation*

---

#### **1. Background and fundamental principles**

This session will establish a foundation for the remainder of the tutorial by delineating the criteria upon which an ontology may be judged. We will first briefly define ontology research and its intersection with computer science. In order to reason upon and draw inferences from data to which an ontology has been applied it is absolutely essential that the relationships be carefully defined, otherwise the data entry is insecure and the results are unpredictable. We will use case studies to illustrate situations to be avoided

and the subtleties of intended meaning in various relationship types. The primary emphasis will be to illustrate key issues through examples so that we may use these to guide ensuing discussions. (1 hours)

## **2. Introduction and survey of existing ontologies**

We will then provide an overview of the relevant biological ontologies that are currently available. This will include all of those included in the Open Biological Ontologies (OBO), the Gene Ontology (GO), the MGED Ontology, the NCI Thesaurus, eVOC, the Plant ontology consortium, the Foundational Model of Anatomy (FMA), Reactome, and the Sequence Ontology (SO). The description will cover the original motivation and scope of the ontology, the relationships used and the logical inferences these support, available data sets to which the ontology has been applied, and the format(s) the ontology is available in. When there is overlap between ontologies we will discuss how this is being resolved. (1.5 hours)

## **3. Ontology creation and critique**

This portion of the tutorial will be spent on techniques for developing high-quality ontologies. Those attending will participate in an interactive process of ontology evaluation during this session, in which we will discuss the relative pros and cons for problematic terms and relationships. The group will learn by seeing ontology creation and evaluation at work. (0.5 hours)

## **4. Conclusion**

In the final section of the tutorial we will assess the outcome of the ontology creation exercise and summarize the lessons it illustrated. (0.5 hours)