Bio-Ontologies: Their Creation and Design

Peter Karp
Bioinformatics Research Group, SRI International, USA
pkarp@ai.sri.com
Carole Goble and Robert Stevens
TAMBIS team, University of Manchester, UK
<carole,stevensr>@cs.man.ac.uk

February 12, 2001

1 Motivation

Ontologies are playing increasingly important roles within a range of bioinformatics applications. A tutorial at ISMB-98 set out to introduce the notion of ontologies, and to motivate for their adoption. That tutorial was followed by ontology workshops at ISMB-98, ISMB-99 and ISMB-2000, for the specialist community, but these have not been for the bioinformatics community as a whole. We propose to follow the introductory tutorial at ISMB-2000 with a more advanced tutorial focused upon the creation and design of bio-ontologies.

The intended audience for this tutorial are (a) developers of bioinformatics databases who wish to have the development of their DB schema guided by ontology methodologies, and (b) developers of bioinformatics ontologies. This tutorial seeks to inform the whole community on how to construct and deliver ontologies within the bioinformatics domain, in the setting of real, practical examples.Attendees should be reasonably familiar with the first half of the material presented at the 2000 ontology tutorial (see
http://img.cs.man.ac.uk/tambis/tutorial/tutorial-final.html.

The University of Manchester and SRI International are leaders in the development and use of bio-ontologies. The two groups represent different approaches and have developed different types of applications. This gives the tutorial the wide experience in developing ontologies needed for the proposed tutorial.

The tutorial will follow the topics and time scale:

- The need for, and nature of, ontologies (30 minutes);
- Knowledge representation for ontologies– the Ontology Inference Layer (OIL) (45 minutes);
- Principles and guidelines for ontology design (45 minutes);
- Ontology development tools (45 minutes).

The tutorial aims to be interactive, with some time set aside for audience participation in the development of a small part of an ontology. This will be centered around the conceptualisation of macro-molecules and their function – features common to many of the current bio-ontologies. The tutorial will be rich in examples taken from the various bio-ontologies; examining the different conceptualisations, encodings and delivery styles and how this is linked to their purpose.

2 The Need For, and Nature of, Ontologies

The tutorial will not take a deep philosophical approach to the nature and use of ontologies. Instead, a practical approach will be adopted, driven by the need to include knowledge of biological and related disciplines within bioinformatics applications. The tutorial will describe the type of knowledge needed, how it can be captured and what must remain within the human domain. Knowledge is couched in terms of a domain's concepts and the relationships held by those concepts. The importance of conceptualisation as a major stage in ontology development will be stressed through the tutorial.

Motivations for ontologies included their use in database interoperation, in machine learning (to provide a generalisation hierarchy to guide the learner), and in development of bioinformatics databases.

3 Knowledge representation for ontologies– the Ontology Inference Layer (OIL)

The tutorial will compare and contrast ontologies with controlled vocabularies, taxonomies, and database schemas. This will highlight the importance in the choice of knowledge representation used to describe the ontology. This more advanced ontology will centre about the different kinds of KR language and what they offer the ontologist.

Various methods of representing knowledge in an ontology will be discussed. A framework will be used to describe these schemes, working along the axes of *expressivity* and *well-foundedness*. The former dimension will be divided into *informal*, *structured* and *formal*. The issues of consistency and interpretation will be discussed. For application developers, the role of an API that hides the representational scheme from the application itself will be explored. The correctness and consistency of each method will be investigated and a brief introduction given to each type of encoding method. References to resources will be given for each type of encoding.

In the 1999 bio-Ontologies workshop Peter Karp presented the XML Ontology Language (XOL). This is an XML based exchange language that can be used to describe the structure of an ontology. Work has progressed on XOL during the past year and the latest incarnation (Ontology Inference Layer: OIL) will be presented during the tutorial.

OIL unites the frame-based modelling style with the reasoning support of description logics (the *structured* and *formal* descriptions used above). It will be shown how OIL can be used to create a range of ontologies from hand-crafted hierarchies of concepts to self-classifying lattices, where the classification is inferred from the description of a concept's properties. The OIL language will be explained and illustrated with the on-line demonstration of its power and versatility using the macromolecule ontology fragment that runs through the tutorial.

4 Principles and guidelines for ontology design

A variety of techniques exist for knowledge representation – acquiring, encoding and delivering a conceptualisation. Irrespective of the knowledge representation used, the stages in building an ontology are much the same. Within the tutorial, a modified version of the well known software engineering V-model will be used to describe the process of building an ontology:

Identify purpose and scope: identifying the intended range of uses of the ontology

Knowledge Acquisition: the process of acquiring domain knowledge from which the ontology will be built

Building the ontology - conceptualisation: identifying the key concepts that exist in the domain, their properties and the relationships that hold between them

Building the ontology - integrating: use or specialise an existing ontology as a foundation

Building the ontology - encoding: representing the conceptualisation in some form of language

Documentation: informal and formal complete definitions

Evaluation: determining the appropriateness of an ontology for its intended application

The ontology fragment of macromolecules and their definitions, functions, cellular locations, etc. will run through this part of the tutorial. The style of question the ontologist must ask him- or her-self during the conceptualisation stage will be richly illustrated using this ontology fragment. It will be shown that different conceptualisations can be used for different ontological tasks and how the properties of the encoding can alter the initial conceptualisation. Examples from each of the three types of encoding will be used to exhibit the different ontologies arising from building the same ontology in different ways.

The tutorial will present principles of ontology design, and will warn the tutee of potential pitfalls that they should avoid. The tutee should gain an overview of the advantages and disadvantages of encoding styles and the costs of each stage of building.

5 Ontology Development Tools

For such complex systems as an ontology, where consistency and clarity are important, the availability of tools that assist the process of conceptualisation and encoding are invaluable. Tools are also important for allowing developers to view their ontology and check development criteria, especially fitness for purpose. The tutorial will survey several existing ontology-development tools, including:

- OilEd, the editor for building OIL based ontologies (University of Manchester and others);
- the GKB Editor (SRI International);
- Protégé 2000(Stanford University).

6 Tutorial Staff

The following three people have developed and will present the tutorial:

Peter Karp: Dr. Peter Karp received the Ph.D. in Computer Science from Stanford University in 1989. He has held positions at the US National Center for Biotechnology Information, at Pangea Systems, and at SRI International, where he directs the Bioinformatics Research Group. He has developed tools for developing ontologies and large knowledge bases, such as the GKB Editor, and has developed a large bioinformatics ontology as part of the EcoCyc project. He has also published several papers on interoperation of bioinformatics databases.

Carole Goble: Carole Goble is a Professor in Computer Science and co-leads the Information Management Group at the University of Manchester. She was/is investigator on a number of projects using ontologies, represented using Description Logics for: medical information systems (PEN&PAD, GALEN), mediating disparate bioinformatics information sources (TAMBIS, TAMBIS-II), and improved protein function prediction using ontologies (Irbane). She is a co-investigator on a basic research project on Description Logic-based ontology servers (CAMELOT). She is currently involved in the development the Ontology Inference layer, which the Information Management Group is developing with Peter Karp and the Free University of Amsterdam.

Robert Stevens: Robert is a bioinformatics researcher at the University of Manchester and has degrees in both Biochemistry and Computer Science. He has experience in the characterisation,

modelling and interoperation between many bioinformatics resources, as well as skills in user requirements analysis and user interface design. He is one of the developers of the TAMBIS bioinformatics source mediation system.