Application of Ontologies in Bioinformatics

Robert Stevens[1], Phillip Lord[2], and Carole Goble[1]

[1] School of Computer Science
University of Manchester
Oxford Road
Manchester
United Kingdom
M13 9pl
Robert.Stevens@Manchester.ac.UK

[2] University of Newcastle Upon tyne
Newcastle Upon tyne
Philip.Lord@Newcastle.ac.uk

Contents

1 Introduction

In this chapter we explore the uses within bioinformatics of some of the ontologies and other ontology-like artefacts, some of which were described in Chapter That chapter provided a motivation for the use of ontology and described the range of ontologies available and the institutional support they now enjoy. In the first edition of this volume we explored why the discipline of bioinformatics has become so interested in the development and use of ontologies **?**. This interest has become consolodated such that development and use of ontologies has become mainstream within bioinformatics. Yet, as we will see in this Chapter, the majority of the use to which ontologies are put within bioinformatics is quite narrow, but the potential uses are wide ranging. we will see that the initial goal of ontology has been basic data integration for use by humans. Ontologies should offer the means to drive computational use of biological data and it is this aspect that we wish to push in this chapter.

The production of data in biology has become industrialised; so must its analysis. The lack of laws captured in some computable form means that much inference in bioinformatics is still reliant on the processing of factual data—the knowledge we have about the entities in the biological world. A common understanding of that which is described by the data collected is obviously a great help in such an endeavour. The primary means of delivering such a common understanding in bioinformatics[3] is by talking about the same entities in the same way—controlling the vocabu-

[3] Here we take a broad definition of bioinformatics to mean the storeage, management and analysis of biological data by computational means to answer biological questions.

lary used to representint suff in data resources. Delivery of controlled vocabulary for a "*de facto* integration" **?** is still the primary use of bio-ontologies.

The need for a common reference for the functional attributes of gene products in the post-genomic era motivated the development of the Gene Ontology (GO) **?**. A common understanding agrees upon thos categories of, for example, molecular function that exists. The labels chosen for those catagories provides a vocabulary (an ontology can deliver a vbocabulary, but it isn't a vocabulary). The control arises from the committment to use that ontology delivered vocabulary to describe the attibutes of classes of gene product across species and community wide resources. As described in Section 3, this has great utility not only in querying resources, but also in their analysis.

Doing a Google search with "`define: ontology`" gives an answer with approximately 20 slightly different definitions. These do, however, cluster into two distinct definitions:

1. A discipline of philosphy concerned with the description of that which exists.
2. A shared understanding of what a community understands about a domain that allows machine reasoning.
3.

In essence, these are both concerned with descriptions of the things in the world. The emaphsis of the second, however, is on the second is that of the *shared* use of the description and its use by computers. As we will see, the idea of labelling and defining 'waht it is to be a member of a ckass; and agreeing the label for that class assits both human and computers in data processing. In a knowledge based discipline such as bioinformatics, having a machine processable form of the knowledge that is used in a wide range of scientific inferences is vital. What we will be able to claim, however, is that description for the sake of description, without including the computer is potentially highly restrictive.

An ontology, according to the philosophers who coined the term, is a description of the categories and membership criteria of that which exist. Computer scientists have latterly taken this term and somewhat loosened its meaning. an ontology still describes things, but the emphasis is on shared understanding of conceptualisations. The goal of a computer science ontology is to enable machines to manipulate symbolic representations of knowledge. Whether or not a broad or narrow view of ontology is taken, ontologies and ontology like artefacts are all about description of the world or human understanding of the world. This can range from an attempt to record the true account of reality through thesaurae, vocabuarlies, classification schems to glossaries. Irrespective of the representation and the level of reasoning supported they all to a greater or lesser extent attempt some description and/or definition of things in the world. Within bioinformatics it is possible to find many knowledge artefacts described as ontologies **?**. In this chapter we will take the broader view of ontology and explore how those descriptions are used within bioinformatics applications.

In section 2 we classify the uses to which ontologies have been put within bioinformatics. Then in Section 3 we look at some case studies of these uses. In Section 4 we discuss the current state and future directions for ontologies within bioinformatics.

2 Classifying Uses of Bio-Ontologies

Ontologies, whether from the computer science or philosophical perspective, are all about description. The applications of ontology within biologyare therefore all rooted in description. Figure 1 shows a classification scheme (very deliberately not an ontology) for the uses of ontology and ontology-like artefacts within biology. Obviously describing the world is a use in itself and conseqeuntly all the uses are *narower* uses of description. Papers from **?** and **?** categorise the potential uses of ontology in to broad categories. These categorisations are still within ours, but we wished to emphasise the role of description, the inter-relatedness of these uses and also to present those uses at a finer granularity. We have already mentioned one of the principle uses of such description in bioinformatics—that of using the labels of the concepts for the delivery of a controlled vocabulary. Other uses of ontologies exploit the structure of the relationships between the concepts. Having annotated data with a controlled bvocabuarly, the structure of the ontology can be used to query instance data, navigate dinstance data, etc. The structure of an ontological description is what many of the uses exploit. to move from a shared understanding for humans to one exploitable by machines ncessitates strict semantics and is further facilitated by the ability of rich expressivity. Such semantic strictness and expressivity does not necessarily afford new uses, but potentially enable more extensive or further uses in the same area.

the uses to which ontological description can be put include, but are not limited to:

Reference Ontology: it is possible, as stated earlier, to regard description of the world as a use in its own right. For the classes of entity within a domain to be defined, it is to be hoped both logically and humanly, then this can be of great utility in its own right. Simply affording a community of discourse an encyclopædia of that which is known acts as a *reference* source for that domain. The Foundational Model of Anatomy **?** can be seen in such a light. Even when there is a lack of consensus about a self-styled reference ontology, it does form an article for discourse. It is easier to argue about definitions than it is to argue about words. For the modeller and others, the act of modelling itself can offer insights. The act of making knowledge explicit can force the asking of questions about assumptions that are all too often implicit in domain discourse.

Controlled Vocabulary: An ontology describes categories of instances in the world or the concepts people use to describe a world. There is a world of instances and humans put these into categories (classes, types, etc.). Humans also decide on labels for those categories and these provide the vocabulary by which humans talk about the categories of instances in the world. Unfortunately, humans decide on lots of different labels for the same categories and often use the same labels for differeent categories. This heterogeneity obviously makes query and analysis

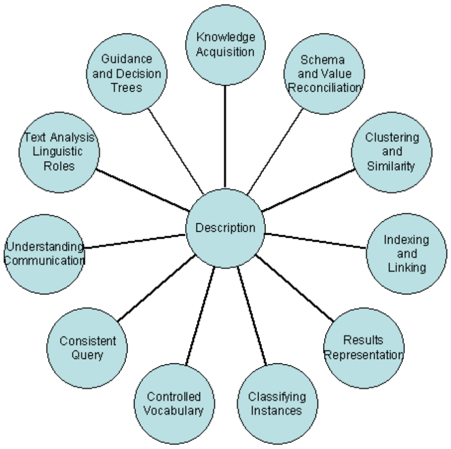
Knowledge Acquisition
Schema and Value Reconciliation
Clustering and Similarity
Indexing and Linking
Results Representation
Classifying Instances
Controlled Vocabulary
Consistent Query
Understanding Communication
Text Analysis Linguistic Roles
Guidance and Decision Trees
Description

Fig. 1. A classification scheme for the uses of ontology and ontology like artefacts within biology.

of data that relies on manipulations of what is known about a biological entity very difficult. By agreeing upon the labels for a category and by committing to use that vocabuarly for the ctegories defined by the ontology then a *controlled vocabulary* has been developed. the development and examples of controlled vocabuaries in biology is described in Section 3.1.

Schema and Value Reconcilliation: Not only do humans disagree on the labels given to categories, but they also disagree on the categories themselves. There are many legitimate ways to describe reality. This can be due to different perspectivbes on the same issues, e.g., taking either a developmental or structural view of anatomy will give different categories. Other descriptions will either be partial or skewed due to some application bias. Many databases exist within bionformatics that represent similary extents or overlapping extents. Thus to get a complete coverage of a domain of interest these data need to be pooled. Unfortuantely differing conceptualisations of the instances the data represent mean that the differing representations need to be reconciled. An agreement by a community on the categories and their definitions—that is,a definition of the extent of the ontology, often expressed either as an intentional definition of the constraints to which an instance must comply or a template of the instances' attributes—means that all the schema and data instances of the client databases have a common model to which they must comply. This use of an ontology to specify a model to drive both schema and value reconciliation is common both within and without bioinformatics. this use is explored in Section 3.2.

Consistent Query: Obviously once there is a common conceptualisation, a common set of labels for the concepts and the instances all comply with that ontology, querying and analysis of data can be greatly eased. given that all resources talk about the same biological entities using the same labels makes querying possible and easy. Different ontological representations aford different kinds of query facility. Simply using a controlled vocabulary allows better querying by exact matching. Using the taxonomic structure of an ontology allows queries to retrieve "instances of this class" which implies all the instances of the sublcasses (as these are instances of the query class). Also, a significant factor in making this possible is doing the annotation or transofrmation of the instances and schema as necessary for the queries. Section 3.3 looks at consistent querying over bioinformatics data resources.

Knowledge Acquisition: Having described the classes of instances in a domain, a practicioner will often want to gather instnaces of those classes. Ontologies can either specify templates for the attributes that instances of a class must be given or describe "what is known about an instance", whether or not it is explicitly stated **?**. Ontologies can be used to generate forms by which instances are gathered or acquired. Similarly, data can be transformed to comply with the ontology to generagte a knowledge base (the combination of ontology and instances of the classes in the ontology). Several examples of these have been seen in bioinformatics and these will be described in Section 3.5. Obviously the ontology then offers the means by which those instances can be queried in a sophisticaed manner.

Clustering and Similarity: Rather than straight forward querying, an ontology can be used to *cluster* data items. For example, if the genes implied by a microaray chip are annotated with Gene Ontology terms, one can take differentially modulated genes and cluster them against the aspects of GO. For instance, the set of upregulated genes on a chip could be clustered about the GO biological process ontology. Taking the lowest common subsumer of those chip members provider the analyst with an idea of what might be happening in the condition under investiation.

This clustering prompts the question of how similar are the members of a cluster? In bioinformatics we are well used to the notion of sequence simlarity and how it is to be interpreted. Recently, as the amount of semantically annotated data has risen, the notion of semantic similarity has become prominent. **?** introduced the notion into bioinformatics and opened the possibility for querying an analyses of data at a semantic level in the style of “these two entities have an 42% functional similarity. The use of description to enable clustering and measures of semantic similairty will be explored in Section 3.6.

Indexing and Linking: As already described, ontologies and ontology like artefacts can provide structured, controlled vocabularies. These are often used to describe data objects. One consequence of this is to *index* those data. Just as with a traditional book index, this is a mechaism for quick retrieval. This has an obvious closeness to querying and searching. Perhaps the most prominent example of indexing the biomedical arena is the use of MeSH (Medical Subject Headings) to index PubMed abstracts. In Section 3.7 we describe the use of knolwedge models in this area and how the linking and navigation task being udnertaken suggests a different form of knowledge model to the formal ontology.

Results Representation:

Classifying Instances: An ontology describes the classes of instances in a domain. Definitions of those classes provide knowledge of how to recognise an domain stance as a member of a particular class. Given a set of facts about instances the ontology can be used to classify those instances to place them into categories or classes.

Understanding Communication:

Text Analysis/Linguistic Roles:

Guidance and Decision Trees: Ontologies, by capturing knowledge about a domain and encapsualting constraints about class membership, can offer guidence around a domain and support decision making processes. In query formulation, for instance, an ontology can inform an application or human opeator information about “this is what iss possible to say about an entity”. In querying about transcription complexes, for instance, an ontology might offer inforamtion about transcription factors, binding sites in promoters, polumerases etc., but not about entities relevant to replication and other possibly irrelvenat processes. The constrains in an ontology can “cut down” the space of possbilities in a large and complex domain such as biology and bioinformatics. Similarly, given a set of facts about symptoms, an ontology can prompt a user to provide more discrimiting facts to distinguish between classes.

There are a range of potential uses for bio-ontologies within bioinformatics. We have presneted a simple classification scheme of their uses in order to help orientation and navigation within the field. In the next section we take examples from biomedicine to illustrate this scheme.

3 Case Studies

3.1 Controlled Vocabulary

3.2 Schema and Value Reconcilliation

3.3 Consistent Query

3.4 Reference Ontology

3.5 Knowledge Acquisition

3.6 Clustering and Similarity

3.7 Indexing and Linking

3.8 Results Representation

3.9 Classifying Instances

3.10 Understanding Communication

3.11 Text Analysis/Linguistic Roles

3.12 Guidance and Decision Trees

4 Discussion

Index