

# Future application of ontologies in e-Bioscience

## Authors

**Marco Roos,**  
Integrative Bioinformatics Unit  
University of Amsterdam  
roos@science.uva.nl

**Scott Marshall,**  
Integrative Bioinformatics Unit  
University of Amsterdam  
marshall@science.uva.nl

**Machiel Jansen,**  
Department of Computer Science  
Vrije Universiteit  
mgjansen@few.vu.nl

**Timo M. Breit**  
Integrative Bioinformatics Unit  
University of Amsterdam  
breit@science.uva.nl

## Abstract

Ontologies can be used for rich annotation of data and the means to share and integrate data. In addition to the annotation of experimental data for search-and-retrieval in an e-bioscience environment, we propose to use semantic modelling to enhance biological problem solving and phenomenon discovery.

## Introduction

Rapid progress in biological research has lead to the availability of large amounts of heterogeneous data from an increasing number of public and commercial resources. To enhance the accessibility and integration of this data, the biology research community has increasingly employed database and web technology. Several standards have been developed for annotation of biological data, including de facto ontological standards such as the Gene Ontology (GO). Following the success of GO, other ontologies are being developed for application to the annotation of biological data.

### *Sharing information*

The Web Ontology Language (OWL) has been accepted by the W3C committee as a standard and is a good candidate for sharing biological information between research communities, users and institutes. In e-science environments, such as the Dutch Virtual Laboratory e-science (VL-e; [www.vl-e.nl](http://www.vl-e.nl)) or the British myGRID ([www.mygrid.org.uk](http://www.mygrid.org.uk)), the annotation of experimental data with the use of ontologies makes several new types of activities possible. For example, we would like the scientist to be able to answer questions such as “Where can I find data relevant to my research?” or, more specifically, “Which experiments have data on protein X?”, or “Which methods from another research area can be of use for my gene expression study?”. Ontologies can provide the vocabulary and the semantics of the annotations needed for these types of queries.

### *Formalising knowledge.*

Ontologies can also be viewed as a formalisation of domain knowledge. Biologists often leave this knowledge largely implicit, but it plays an essential role in biological research and must be made explicit for use by computational biology applications. Ontologies can assist the biologist not only in the creation of semi-formal, qualitative models associated with biological problems, but also in the formulation of hypotheses and conclusions. This allows for the application of ‘knowledge-intensive’ methods such as automatic refinement of existing models (e.g. Shrager *et al.*, 2002), as well as reasoning about scientific experiments. We believe that such semantic modelling using ontologies makes new options available for exploration by researchers, especially for complex biological problems. In our view, factual biological knowledge and the procedures and methods in experimental research can all be captured in an ontological form. In this manner we strive for a knowledge-intensive processing of biological data as opposed to a direct, ‘knowledge-lean’ approach, which is currently prominent in bioinformatics. We believe that semantic modelling enables one to automate some important tasks involved in biological problem solving and phenomenon discovery.

Taken together, these developments increase the usefulness of e-science for biological research (‘e-bioscience’).

## **Approach**

What we propose can be regarded as a two-pronged approach based on the two intended applications. First, we advocate the use of ontologies for the annotation of data. In the context of biology, this means using standardized vocabularies such as the GO in combination with Life Sciences Identifiers (LSID). They provide the basis for metadata-related services and data integration tools.

Second, we would like to use ontologies to facilitate the building of qualitative biological models subject to rigour and logic, which could then be used as the basis for knowledge-intensive methods. Qualitative models could help with hypothesis formation, model-refinement, and phenomenon discovery. In this respect, comparison of models can be a powerful instrument as long as the models have a common ontological commitment (hence the importance of using standards). Furthermore, if we have explicit qualitative models, we are able to reason about them. It will be interesting to see if current ontologies suffice for application in our first case studies.

## **Example**

Our example uses analogy between models to provide the scientist with hypotheses. Imagine a human disease that is caused by a certain gene. Information about the human variant of this gene, its annotations, is present in a human genome database. Using its LSID, the zebrafish variant of this gene can be found in the genome database of zebrafish, which is annotated using the same system. Because zebrafish can more easily be experimented on than humans, the amount of knowledge about the zebrafish variant of the gene far exceeds that of the human variant. Using a

knowledge-intensive method, we can use the model in zebrafish to suggest a likely model (or model extension) for the human case. This model can then be tested in light of (possibly scarce) human experimental data. To make it a bit more interesting, we could assume that the exact gene is not present in zebrafish. However, if information associated with our human gene, as provided by our knowledge model, can be found in zebrafish, a hypothesis can be suggested for the human case based on zebrafish knowledge. [Common functionality, not necessarily using the same genes, is not uncommon in nature.]

## Risks and challenges

### *Lack of formalised biological knowledge*

The steps performed during experimental research, the (qualitative) models used, and the impact of experimental results on hypotheses and existing beliefs, are seldom formalised and often largely left implicit. Much of the knowledge used can only be obtained from experts. This also applies to knowledge that is used by researchers to design their next experiment in an experimental series. In fact, a semi-formal, semantic characterization in the form of a transcribed model is very unusual. Biological research papers traditionally offer diagrammatic models and text, and more recently references to experimental results and resources on the web. We can only hope experimental biologists to formalise these steps for e-scientists, if they start to appreciate the value of e-bioscience for their own research. However, we can make some positive observations as well. First of all, we have no reason to believe that our approach can only be successful if all knowledge is available. To the contrary, it has been shown that incomplete models can produce useful results (e.g. Chrisman *et al.*, 2003; Shrager *et al.*, 2002). Our aim is to keep the initial models as simple as possible. Furthermore, as in the case of learning ontologies, methods such as text-mining can help in creating knowledge models (semi) automatically.

### *Lowering the barrier to annotation*

Most scientists do not want to spend large amounts of time adding annotations to their hard-earned data for e-scientists. It is up to the e-science community to lower the barrier for annotation. For example, in a robust implementation, it would not be possible to enter non-existent identifiers for experimental entities. We note here the difference between the annotation of single data elements such as genes or proteins and the annotation of an entire experiment. The entry of annotation for an experiment can potentially require complex structures (e.g. to represent a series of experiments and their parameters, each with a possibly associated hypothesis or conclusion). It remains a challenge to make the process of annotation robust and intuitive.

### *Ontology construction*

For creating knowledge models it is known that the required ontologies are incomplete or do not yet exist. Ontology construction is such a time-consuming and laborious process that there has been a growing interest in the automatic acquisition of concepts, relations and rules from various forms of data. Techniques from machine learning, natural language processing, information retrieval, question answering and data mining can be used to automate ontology construction. Domain experts, biologists in this case, may question the validity of the constructed ontologies. Indeed, the resulting ontologies are often incomplete and may contain flaws, but their construction cost in terms of time and effort is low. Furthermore, knowledge

acquisition techniques (i.e. extracting knowledge from human experts) can be used to verify or improve the content of automatically constructed ontologies quite rapidly. The net effect is a quicker construction and curation of valid ontologies, while consuming less time of experts.

## Conclusions

It is our position that e-bioscience can benefit from the use of ontologies in several ways, such as sharing information, annotation and retrieval of data. In addition, ontologies can be applied to the formalisation of existing domain knowledge and therefore used in hypothesis formation, knowledge modelling, and phenomenon discovery. Experiments can be associated ('annotated') with knowledge models, for example, a hypothesis that was tested in the experiment. This brings the two applications of ontologies together: data annotation and knowledge formalisation. A complicating factor is that although biologists use such knowledge extensively, little is formalised for computational purposes. We realize that ontology construction and subsequent annotation is a time-consuming effort, which can benefit from automatic methods. In the initial phase, we do not expect our approach to do much better than more traditional 'direct' approaches. However, it is our conviction that ontology use will pay off in the long run, because formalised knowledge can be readily reused and integrated to tackle larger biological research challenges. This line of e-bioscience research has only just begun and will provide a technology-push for computer-based experimental biology. It is a promising prospect that e-bioscience will efficiently increase our understanding of biological systems through a cycle of data annotation, knowledge modelling, and computational experiments.

## Acknowledgements

We would like to acknowledge the critical reading and suggestions of Han Rauwerda.

## References

- Chrisman, L., P. Langley, S. Bay, and A. Pohorille.** 2003. Incorporating biological knowledge into evaluation of causal regulatory hypotheses. *Pacific Symposium Biocomputing*: 128-139.
- Shrager, J., P. Langley, and A. Pohorille.** 2002. Guiding revision of regulatory models with expression data. *Pacific Symposium Biocomputing*: 486-497.