BIO-ONTOLOGIES: A KNOWLEDGE REPRESENTATION RESOURCE IN BIOINFORMATICS

Carmen Galvez
University of Granada
Granada, Spain
cgalvez@ugr.es

Abstract

Bioinformatics manages the information that has been gathered in databases since the advent of the molecular biology technological revolution. The successful research is based in interpretations of that information that can be accessed and managed computationally, which is a difficult task. An attempt to solve that problem is to use ontologies. Ontologies are computational formalisations of the knowledge about a given domain, allowing computers to manage the information in a semantic level. In medical informatics, ontologies have been used for a longer period of time to produce controlled lexicons for coding schemes. Bio-ontologies define the basic terms and relations in biological domains and are being used among others, as community reference, as the basis for interoperability between systems, and for search, integration and exchange of biological data. The most successful ontologies applied in Bioinformatics are the ones in the *Open Biomedical Ontologies (*OBO) project. At the same time, the *Web Ontology Language* (OWL) is a official proposal for ontologies implementation in the semantic web. In this article, we review the current position in bio-ontologies. We review this trend and what benefits it might bring to ontologies and their use within biomedicine**.**

Keywords

Ontologies, Bio-ontologies, Knowledge Representation, Open Biomedical Ontologies, Gene Ontology

1. INTRODUCTION

In biology, ontologies *allow scientists to specify to any degree of resolution, how data, terminology (i.e. controlled vocabularies) concepts and ideas all relate to each other* [1]. Ontologies play a pivotal role in the semantic web as vehicles for knowledge representation. Many ontologies have already been developed and are used in several areas, including bioinformatics and systems biology. They are considered to be an important technology for the semantic web. They are used for communication between people and organizations by providing a common terminology over a domain. Besides, ontologies provide the basis for interoperability between systems. They can be used for making the content in information sources explicit and serve as an index to a repository of information. Further, they can be used as a basis for integration of information sources and as a query model for information sources. They are used for different functions, such as web agents [2] and web services [3], GRID technology [4] or data-mining and text-mining [5, 6].

Although ontologies have been around for a while, it is only during the last decade that the creation and use of biological ontologies have emerged as important topics. The work on biological ontologies is now recognized as essential in some of the grand challenges of genomics research [7] and there is much international research cooperation for the development of biological ontologies (e.g. *Open Biomedical Ontologies* (OBO) and the use of biological ontologies for the Semantic Web. The number of researchers working on methods and tools for supporting ontology engineering is constantly growing and more and more researchers and companies use ontologies in their daily work. The use of biological ontologies has grown drastically since database builders concerned with developing systems for different (model) organisms joined to create the Gene Ontology (GO) Consortium in 1998 [8]. The goal of GO was and still is to produce a structured, precisely defined, common and dynamic controlled vocabulary that describes the roles of genes and proteins in all organisms. Another milestone was the start of Open Biomedical Ontologies as an umbrella Web address for ontologies for use within the genomics and proteomics domains [9]. The member ontologies are required to be open, to be written in a common syntax, to be orthogonal to each other, to share a unique identifier space and to include textual definitions. Further, in systems biology ontologies are used more and more, for instance, in the definition of standards for representation and exchange of molecular interaction data.

2. BIO-ONTOLOGIES

The aim of ontologies in biology is to express the complex knowledge related to biology in a way that is computationally tractable. There are many biological ontologies. They differ in the type of biological knowledge they describe, their intended use, the level of abstraction and the knowledge representation language. There are ontologies focusing on things such as protein functions, organism development, anatomy and pathways. Most biological ontologies are controlled vocabularies, taxonomies or thesauri, but there are also ontologies that are knowledge bases and use ***Web Ontology Language* (**OWL), and **Resource Description Framework** (RDF) Schema as their representation language. With respect to the abstraction level the ontologies may range from high level ontologies that define general biological knowledge to ontologies that describe selected aspects.

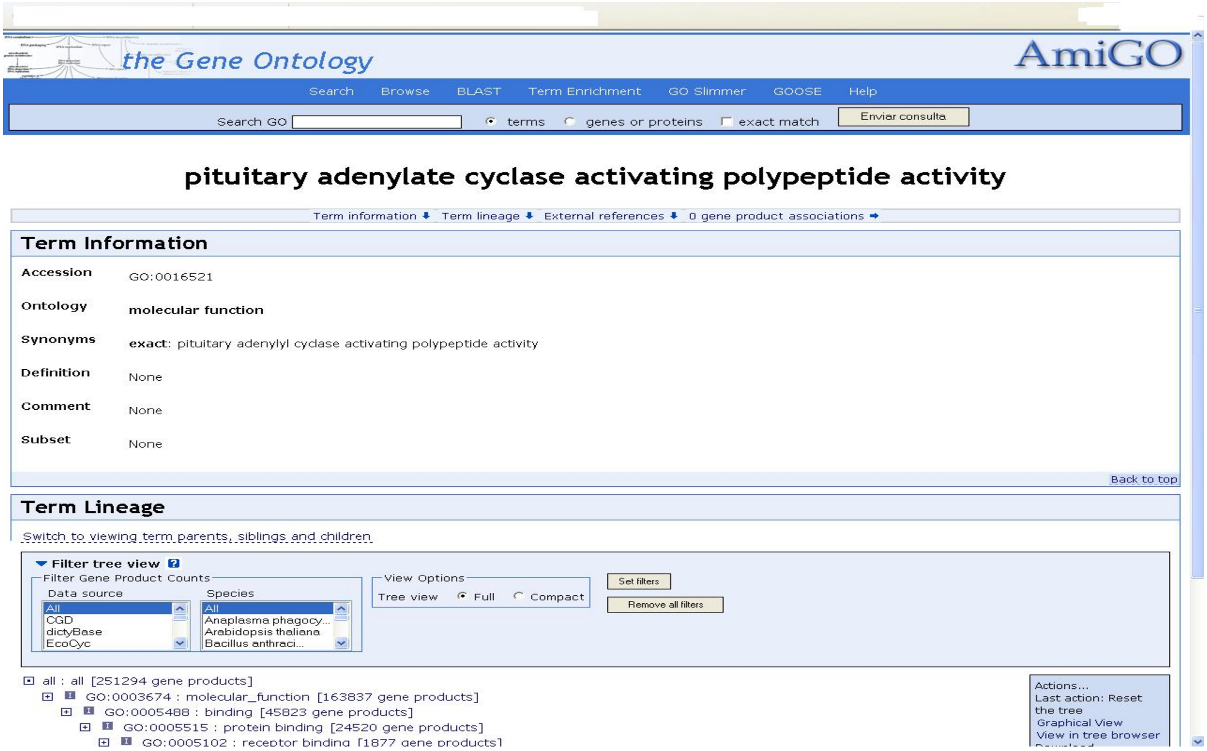
The variety of ontology-like structures will range from controlled vocabularies, thesauri, structured controlled vocabularies, directed acyclic graphs, frame-based systems, up to rich logical axiomatization of our knowledge [10]. The use of the word ontology within biology is quite recent. The Molecular Biology Ontology (MBO) [11] was an early attempt to begin to define the entities in the domain to promote consistent interpretation across resources. A second phase saw the adoption of ontology by the biological community itself. Preeminent amongst these is the Gene Ontology (GO) [12].

The GO [13, 14] provides an ontology that describes attributes of the gene products of an abstract cell. GO offers a way of dealing with the semantic heterogeneity of gene product annotations in different databases: the annotations on different databases point to the same GO term. The Gene Ontology is responsibility of the GO consortium, a joint project formed by different organism databases that was started by FlyBase [15], Mouse Genome Informatics (MGI) [16] and the *Saccharomyces* Genome Database (SGD) [17]. The main component of GO are the terms and the relationships that connect those terms. GO is divided in three independent ontologies: *molecular function*, *biological process* and *cellular component*.

- *The molecular function* describes basic and concrete molecular roles of gene products (*e.g. thioredoxin-disulfide reductase activity GO:0004791*).
- The *biological process* is made of different molecular functions and it describes a higher level role (*e.g. development GO:0007275*).
- The *cellular component* ontology represents the structure of eucaryotic cells (*e.g. organelle GO:0043226*).

Together these capture three of the major aspects that biologists wish to describe about the gene products they place in databases. As genome database providers commit to the GO (that is, they agree with its view of the world) and adopt the terminology delivered by the GO, then each resource describes its gene products in a common form. This sharing, together with the structure provided by the relationships between terms in the GO makes querying of within and between resources possible. The whole ontology is implemented using *Directed Acyclic Graphs* (DAGs): multiple parent-child relationships are allowed in the structure, but cycles (a term being a child of itself) are prohibited. The *top* of the hierarchy is populated by general terms and as we move *deeper* (more terms in the path) the terms become more specialised. The terms on the edge of the path are called *leaves* and terms in the path itself are called *nodes*.

GO can be explored using various tools, the most common one being the AmiGO web interface (Figure 1). GO ontologies can be obtained in different ways, including OBO format, flat files, XML, MySQL tables, etc. Annotations of other databases to GO are available in a list. The databases that include GO annotations are: SGD (*Saccharomyces cerevisiae*), FlyBase (*Drosophila melanogaster*), TAIR (*Arabidopsis thaliana*), WormBase (*Caenorhabditis elegans*), RGD (*Rattus norvegicus*), Gramene (*Oryza sativa*), ZFIN (*Danio rerio*), DictyBase (*Dictyostelium discoideum*), TIGR, Sanger GeneDB, GenBank and UniProt.



pituitary adenylate cyclase activating polypeptide activity

Term Information

Accession	GO:0016521
Ontology	molecular function
Synonyms	exact: pituitary adenylyl cyclase activating polypeptide activity
Definition	None
Comment	None
Subset	None

Term Lineage

Switch to viewing term parents, siblings and children

Filter tree view

Filter Gene Product Counts

Data source: All, CGD, dictyBase, EcoCyc

Species: All, Anaplasma phagocy..., Arabidopsis thaliana, Bacillus anthraci...

View Options

Tree view ● Full ○ Compact

```
all : all [251294 gene products]
  GO:0003674 : molecular_function [163837 gene products]
    GO:0005488 : binding [45823 gene products]
      GO:0005515 : protein binding [24520 gene products]
        GO:0005102 : receptor binding [1877 gene products]
```

Actions...
Last action: Reset the tree
Graphical View
View in tree browser

Figure 1. Representation of the molecular function "pituitary adenylate cyclase activating polypeptide activity" in the Gene Ontology

3. FORMALISING KNOWLEDGE IN ONTOLOGIES

Ontologies can be seen as defining the basic terms and relations of a domain of interest, as well as the rules for combining these terms. Ontologies are models that represent knowledge about a domain in a computable way. Ontologies differ regarding the kind of information they can represent. *Concepts* represent sets or classes of entities in a domain. The concepts may be organized in taxonomies, often based on the *is-a* relation or the *part-of* relation. *Instances* represent the actual entities. They are, however, often not represented in ontologies. Further, there are many types *of relations.* For instance, one type is the group of taxonomic relations such as the specialization relationships. Finally, *axioms* represent facts that are always true in the topic area of the ontology. Ontologies can be classified according to the components and the information regarding the components they contain. A simple type of ontology is the *controlled vocabulary.* These are essentially lists of concepts. When these concepts are organized in an *is-a* hierarchy, we obtain a *taxonomy.* A slightly more complex kind of ontology is the *thesaurus.* In this case the concepts are organized in a graph. The arcs in the graph represent a fixed set of relations, such as synonym, narrower term, broader term, similar term. The data models allow for defining a hierarchy of classes (concepts), attributes (properties of the entities belonging to the classes, functional relations), relations and a limited form of axioms. The *knowledge bases* are often based on a logic.

In the Semantic Web ontologies are the mechanism for providing a vocabulary that will describe data held in a common data model. The Semantic Web is a means to build a World Wide Web where the semantic content is accessible for computers, not just for the human users. One of the main components of the Semantic Web are *ontologies.* The vocabulary and the semantics provided by the ontology all facilitate machine processing. Ontologies are usually collections of classes, each class being a group of individuals, where the classes are linked by different logical relationships, creating a structure. One of the main aims for an ontology is to create a shared understanding. This shared understanding can be extended to computers. Whilst not having the same *understanding* as a human, the computer can make inferences about the symbols themselves. By enabling a computer to do more sophisticated processing, it is possible to gain more added value from the process of annotating data with terms from an ontology.

Ontologies can be produced in different *Knowledge Representation* (KR) languages. An ontology and its components can be represented in a spectrum of representation formalisms ranging from very informal to strictly formal [18]. In general, the more formal the used representation language, the less ambiguity there is in the ontology. Formal languages are also more likely to implement correct functionality. In the informal languages the ontology content is hard-wired in the application. This is not the case for the formal languages as they have a well defined semantics. However, building ontologies using formal languages is not an easy task. These languages differ in their expressivity: the more expressive a language is the more complex can be the knowledge represented by the ontology. During the initial development of Semantic Web technologies there has been an evolution from data exchange standards like XML (*eXtensible Markup Language*) to ex-change languages with more semantics like RDF (*Resource Description Framework*). OWL (*Web Ontology Language*) [19] is the next layer in semantic expressivity ahead of RDF [20]. OWL is a W3C official proposal for a semantic exchange language in the Semantic Web.

On the other hand, biomedical terminologies are typically large, covering tens to hundreds of thousands of entities. Until recently, no widely used ontology development environments (as opposed to ontology editors, to take a software development analogy) were available and ontologies were developed essentially “by hand”, or with rudimentary tools such as file system-like tree editors. In the past fifteen years, Protégé has emerged as the leading ontology editor across disciplines. At the same time, description logics (DL) have superseded frame-based languages to become the leading formalism for representing ontologies. Finally, Semantic Web technologies are playing an increasing role in knowledge representation. This cross-discipline view is in contrast to that in bioinformatics and medical informatics. Within bio-ontology, in-house tools have been developed by the Gene Ontology Consortium in the form of DAG-Edit and latterly OBO-Edit. Medical informatics has used a variety of tools, either proprietary or open-source. In this section we briefly review some knowledge representations and ontology development tools.

3.1 Protégé

Developed by the Stanford Medical Informatics group with funding from various US Government agencies in the past fifteen years (and now a core technology of the National Center for Biomedical Ontology), Protégé is the leading ontology editor across disciplines, with a community of about 50,000 users, representing research and industrial projects in more than 100 countries. Originally developed for representing frame-based ontologies, Protégé has evolved, in collaboration with the University of Manchester, to represent ontologies in the OWL, based on description logics. Many large biomedical ontologies have adopted Protégé for their representation, including the Foundational Model of Anatomy (frame-based) and the NCI Thesaurus (DL-based), though Protégé is not used for the majority of OBO ontologies. Beside the support of OWL, recent changes for Protégé include support for exporting Protégé ontologies into a variety of formats (e.g., RDF/S, OWL and XML Schema).

3.2 Semantic Web technologies

In addition to contributing to specialized domains such as health care and life sciences, the World *Wide Web Consortium* (W3C) creates the very infrastructure of the Semantic Web. The W3C originally developed the specifications of HTML, the markup language used to represent documents in the World Wide Web. Similarly, the W3C produced the specifications of other formalisms for representing documents, resources and ontologies, including XML, RDF/S, OWL. Collectively know as Semantic Web technologies, these specifications define the building blocks of the Semantic Web. Building upon them, additional formalisms are defined to represent, for example, rules. Some of these technologies will be briefly reviewed, with emphasis on their relations to biomedical applications.

The RDF extends the capabilities of the extensible markup language XML as it enables many-to-many relationships between resources and data. The resulting structure is a graph in which the nodes are resources (identified by a Uniform Resource Identifier or URI) or data (e.g., strings, numerals) and the edges are relationships (called properties). RDF integrates limited inference rules, enabling for example to define subclasses and sub-properties. Some extensive resources such as UniProt have already been converted to RDF. The BioRDF task force of the W3C Semantic Web Health Care and Life Sciences Interest Group currently investigates methods whereby existing resources can be converted to RDF. The OWL plays a central role in bio-ontologies and was mentioned multiple times already. OWL DL, the description logic flavor of OWL, is particularly well suited for representing bio-ontologies. The inference supported by RDF and OWL is limited compared to rule-based languages. The role of ontologies in this context is to provide the vocabulary used in the rules.

4. APPLICATIONS AND LIMITATIONS OF BIO-ONTOLOGIES

Regarding biological ontologies the main focus has been on data source annotation, ontology-based search, data integration, data exchange and the use of ontologies as a community reference. Many biological data sources use ontologies for annotation of their data entries and many tools exist to support annotating data sources or to predict annotations for data entries. The annotations are used in several ways. Search engines can take advantage of the annotations as they give extra information. Further, several kinds of systems use GO annotations to compute a semantic similarity measure between entries in data sources. Entries annotated with similar sets of GO terms are considered likely to be similar themselves [21]. Such a similarity measure can be used for data integration and grouping of data entries [22]. There are also many tools that use GO annotations to interpret gene expression analysis on multiple genes. For instance, given a list of genes from a microarray experiment, systems calculate over- or underrepresentation statistics for each GO term related to the genes in the experiment. This provides a description of significant features of the genes in the list. Ontologies and annotations are also used in text-mining. Ontologies are also used in different steps in ontology-based search. An ontology can be used as an index to the information in the information sources. A user can browse the ontology and use the terms in the ontology as query terms. For instance, TAIR Keyword Browser (Fly), GO Fish (Yeast, Fly, Mouse, Worm) and MGI GO Browser (Mouse) use GO to browse databases. MeSH is used to index PubMed, an archive for biomedical and life sciences journal literature, and GO PubMed connects GO to PubMed.

The use of ontologies can help to overcome interoperability problems. In order to achieve interoperability many ontology-based approaches to the information integration have been developed in different fields. As for the bioinfomatics this problem still remains open. From one side biologist needs to have a possibility to analyze a wide range of data, to pose complex queries over different resources [23]. From the other side, existing biological databases are encoded in different and incompatible formats; they have different data models, from flat-files to object-oriented databases. There are also no naming conventions between databases. At the same time, there are only a few reusable bio-ontologies. This is partially because of the diversity of their representation forms, because of the explicitness of their semantics and the range applications they address. Moreover, still there are also no approaches for integration of bio-ontologies. However, it is evident that when developing a new application for the integration of biological data for different tasks, the bio-ontology put in the base of such an application should not be designed from scratch, rather it should integrate all or some modules of existing bio-ontologies, since the process of bio-ontology building is a high-cost process. All this requires a very close collaborative work of people from biology and computer science community.

5. CONCLUSIONS

With the recent advance in bio-technology and bioinformatics, numerous genome databases from various biological communities have been developed to assist in genomic research. There are hundreds of genomic and biological databases open for common access throughout the World Wide Web. When using more than one data store or analysis tool, a biologist needs to be sure that the knowledge within one resource can be reliably compared to those in another. Information integration in general and in biology in particular requires a consistent shared understanding of the meaning of that information. Bio-ontologies provide a shared and common structure of a domain thus giving a common understanding of this domain, and may be used for overcoming semantic heterogeneity. However, for bio-ontologies there are also the same problems that exist for ontologies in general, namely the creation of ontology development tools (editors), development of methodologies supporting the development and use of ontologies.

References

[1] Neumann EK, Miller E, Wilbanks J. What the Semantic Web could do for Life Sciences. Biosilico 2004; 2(6): 228–236.

[2] Hendler J. Agents and the semantic web. IEEE Intelligent Systems Journal 2001; 16:30–37.

[3] Nicholas Gibbins, Stephen Harris, and Nigel Shadbolt. Agent-based semantic web services. Journal of Web Semantics, 1(1): 141–154, 2004.

[4] Stevens RD, Robinson AJ, Goble CA. MyGrid: personalised bioinformatics on the information grid. Bioinformatics 2003;19: i302–i304.

[5] Li Y, Zhong N. Web mining model and its applications for information gathering. Knowledge-Based Systems 2004; 17: 207–217.

[6] Kim J-D, Ohta T, Tateisi Y, Tsujii J. GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics 2003; 19: i180– i182.

[7] Collins F, Green E, Guttmacher A, Guyer M. A vision for the future of genomics research. Nature 2003; 422: 835-847.

[8] GO, The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 2000; 25(l): 25–29.

[9] OBO, Open Biomedical Ontologies; http://obo.sourceforge.net/.

[10] Altman R, et al. RiboWeb: An ontology-based system for collaborative molecular biology. IEEE Intelligent Systems 1999; 14(5): 68–76.

[11] Bada M., et al. A short study on the success of the Gene Ontology. Journal of Web Semantics 2004; 1: 235–240.

[12] Mungall, CJ. Obol: integrating language and meaning in bio-ontologies. Comparative and Functional Genomics 2004; 5(6-7): 509-520.

[13] Lewis S. Gene Ontology: looking backwards and forwards. Genome Biology 2005; 6:103.

[14] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 2000; 23:25–29.

[15] The FlyBase Consortium. The FlyBase database of the drosophila genome projects and community literature. Nucleic Acid Research 1999; 27:85–88.

[16] Blake JA. The mouse genome database (MGD): expanding genetic and genomic resources for the laboratory mouse. Nucleic Acid Research 2000; 28:108–111.

[17] Christie KR, et al. Saccharomyces Genome Database (GSD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related se-quences from other organisms. Nucleic Acid Research 2004; 32:D311–D314.

[18] Jasper R, Uschold MA. Framework for understanding and classifying ontology applications. In: Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends 1999.

[19] Grigoris A, Harmelen, FV. Web Ontology Language: OWL. In: Handbook on ontologies (International Handbooks on Information Systems) 2004: 67–92.

[20] Wang X, Gorlitsky R, Almeida JS. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nature Biotechnology 2005; 23(9): 1099–1103.

[21] Lord P, Stevens R, Brass A, Goble C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003; 19(10): 1275-1283.

[22] Jakoniene V, Rundqvist D, Lambrix P. A method for similarity-based grouping of biological data. In: Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences 2006; LNBI 4075: 136–151.

[23] Gerstein M. Integrative database analysis in structural genomics. Nature Structural Biology 2000; 7 Suppl: 960-963.