Chapter 15
Ontologies in Biology

Janet Kelso, Robert Hoehndorf, and Kay Prüfer

15.1 Introduction

Modern biology is a data-producing, data-driven science. Biological databases covering the domains of sequence, structure, phenotype, and many other types of biological information are core resources for biomedical research. Recent advances in molecular biology, coupled with rapid development of high-throughout technologies, have lead to the exponential growth of databases housing information about the sequences, functions and localizations of genes and proteins for a wide range of organisms. The bottleneck is therefore no longer the production of data, but the integration and analysis of this data. In order to make biologically meaningful discoveries, researchers require the ability to query and extract the biological information available from a variety of sources, and to integrate this information in meaningful ways. However, there are a number of obstacles that make such integrated analyses difficult.

(i) With the exception of the major nucleotide and protein databases biological databases are generally developed and maintained by the community of scientists that are interested in the scientific questions that can be addressed by the data being stored in the database. As such it is common that biological data is stored in geographically disparate locations, using different technologies and representations. Redundancy or partial overlap in stored data is common.
(ii) The integration of biological data has been severely hindered by ambiguities in terminology, semantics and storage. Synonyms and abbreviations are widely used and often applied with conflicting meanings. Homonyms are present both within and between biological sub-disciplines. Further, the definitions of

J. Kelso (✉)
Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany
e-mail: kelso@eva.mpg.de

R. Hoehndorf (✉)
Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany;
Department of Computer Science, University of Leipzig, Leipzig, Germany
e-mail: leechuck@leechuck.de

R. Poli et al. (eds.), *Theory and Applications of Ontology: Computer Applications*,
DOI 10.1007/978-90-481-8847-5_15,

> even fundamental biological concepts such as *organism*, *gene* and *species* are not universally agreed upon. It is likely that these problems have arisen as a side-effect of our constantly evolving understanding of biological systems, and as a result of the gradual merging of historically distinct sub-disciplines as biological research becomes more integrative.

In 1998 Steffen Schulze-Kremer presented a paper at the Pacific Symposium of Biocomputing (Schulze-Kremer, 1998) in which he discussed the application and potential future applications of ontologies in molecular biology. Both in this paper, and in a later paper (Schulze-Kremer, 2002), he clearly identified the information exchange and data integration problems prevalent in the biological sciences saying: "Many researchers and databases use (at least partially) idiosyncratic terms and concepts for representing biological information. Often, terms and definitions differ between groups, with different groups not infrequently using identical terms with different meanings. The concept 'gene', for example, is used with different semantics by the major international genomic databases." He proposed ontologies as a means to provide standardized nomenclature for the rapidly growing databases of sequence, structure, expression, metabolic and regulatory data for many organisms.

Recent years have seen a growing trend towards the development and adoption of ontologies for the management of biological knowledge. Ontologies and controlled vocabularies for various domains of the biomedical sciences have been developed, largely in an effort to provide a shared language for communicating biological information. Ontologies are viewed by the biomedical community as a powerful means to represent, analyze and integrate biological information.

More historically, however, much of the original basis of biology is in the classification of domains. An early example of the classification of organisms are the taxonomies formulated by Linnaeus. The controlled vocabulary of MeSH terms used by Entrez at the NCBI portal[1] of the National Library of Medicine are another example of where a structured set of terms are used to classify publications and index them for searching.

From a biologist's perspective, a controlled terminology with structured relationships is useful in many domains. It provides a consistent and defined nomenclature and provides structured access to possible terms and relationships.

The major recent utilisation of ontologies in biomedicine has been largely to provide a common terminology for a variety of domains (discussed later in this chapter). Successful utilisation of ontologies is dependent upon multiple factors including their usability, design and on their broad adoption by the community. There has been some debate over whether a single all-encompassing ontology or smaller domain or task-specific ontologies are more useful.[2] Smaller ontologies take less time to build and are simpler to maintain and grow. As a result of their practicality, smaller ontologies relevant to distinct domains of molecular biology have been rapidly developed and put to use. In order for ontologies in a domain to be

[1] http://www.ncbi.nlm.nih.gov/

[2] http://en.wikipedia.org/wiki/Upper_ontology_(computer_science)

accepted as the standard, community involvement and adoption are essential and this community agreement has been a hallmark of the development of the modern "bio-ontologies". Many of the widely used and accepted ontologies have been built by consortia, and are designed with specific applications in mind. The most common application of these bio-ontologies has been the formalisation of domains of biomedical knowledge through explicit and unambiguous definition of terms used for the description of biological data. This has been achieved through the naming and definition of relevant entities within biomedical domains, and the specification of the relationships that exist between them. Additionally, ontologies that specify the schemas of knowledge bases have also been valuable in providing a basis for a variety of standards specifications for the collection of gene expression (Whetzel et al., 2006), and sequence data (Field et al., 2006).

In this chapter we will explore a few of the modern bio-ontologies and will discuss their scope, strengths and weaknesses. We will include a short description of the resources and applications that have been developed around these bio-ontologies with a view to showing how valuable these ontologies have become in supporting biological research. The application of formal ontological principles to the design of many of the biological ontologies has lagged behind the development of "light-weight" domain ontologies, and as a result the scope of applications remains restricted. We will discuss some of the criticism of the lack of formality in the bio-ontologies, and provide some ideas about how formal logics can be used to address the growing need for ontology integration.

15.2 Ontologies in Biomedicine

The use of ontology in biomedicine has a long history. Some of the older medical terminological systems are discussed in some detail in Chapter 16 by Herre, this volume. We will not discuss these in any detail, but will focus on the second generation of biomedical ontologies which appeared in recent years.

Growth in the development and use of ontologies in biology in the last 10 years has been driven by the need for biologists to organise large volumes of data being generated in molecular biology. To share this data effectively it was necessary to identify and agree on the relevant concepts and select a shared set of terms for the description of these domains. Based on this early start a large and growing number of bio-ontologies have arisen.

15.2.1 The Open Biomedical Ontologies

The Open Biomedical Ontologies (OBO) project is an umbrella organization which hosts a library of ontologies for the biomedical domain.

To be included in the OBO, ontologies need to conform to a set of criteria design which ensure their quality and inter-operability. The OBO co-ordinators provide guidelines for ontology development and facilitate communication between the ontology developers in order to support the development of such ontologies.

The OBO Foundry,[3] a project attempting to increase the formal rigour of the OBO ontologies through the application ontological principles, is based on the following set of principles:

- The ontology must be *open* and available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed under the original name or with the same identifiers.

  Making the ontologies freely available is intended to increase acceptance and use of the ontology, which in turn ensures that the content is accurate and reflects the views of the community.
- The ontology is in, or can be expressed in, a common shared syntax. This may be either the OBO syntax, extensions of this syntax, or OWL.

  The motivation for this principle is that it aids in facilitating inter-operability and permits the development of tools and methods which can then be usefully applied to multiple domains.
- The ontology possesses a unique identifier space within the OBO Foundry.
- The ontology provider has procedures for identifying distinct successive versions.
- The ontology has a clearly specified and clearly delineated content. The ontology must be orthogonal to other ontologies already lodged within OBO.

  The major reason for this principle is to allow two different ontologies, for example anatomy and process, to be combined through additional relationships. These relationships could then be used to constrain when terms could be jointly applied to describe complementary (but distinguishable) perspectives on the same biological or medical entity. As a corollary to this, the OBO Foundry strives for community acceptance of a single ontology for one domain, rather than encouraging rivalry between ontologies.
- The ontologies include textual definitions for all terms.
- The ontology uses relations which are unambiguously defined following the pattern of definitions laid down in the OBO Relation Ontology.
- The ontology is well documented.
- The ontology has a plurality of independent users.
- The ontology will be developed collaboratively with other OBO Foundry members.

A wide variety of biomedical domains are covered by the OBO including ontologies of anatomy, development and disease for a number of key organisms, ontologies of biological sequence, function and process, and ontologies of biochemistry, cell types and behaviour.

Here we discuss the Gene Ontology as an example of a successful and widely used biomedical ontology which forms part of the Open Biomedical Ontology Foundry collection.[4] More than 65 additional ontologies in various domains and stages of development are included in the OBO Foundry. Some of the key ontologies are described in Table 15.1.

[3] http://www.obofoundry.org/

[4] http://www.obofoundry.org/

Table 15.1 Ontologies that form part of the OBO Foundry. There are many more ontologies that are part of the OBO library. These cover a broad range of domains including the anatomies and development of various eukaryotic organisms (plant and animal), disease, behaviour

Domain	Ontology name		Description
Molecular biology	Gene ontology	GO	Three ontologies which describe the molecular function, cellular location and biological process for genes and gene products
Molecular biology	Sequence ontology	SO	Ontology for biological sequence annotation and the description of sequence objects in databases
Organism anatomy and cell type	Common anatomy reference ontology	CARO	A common template for anatomical ontologies
Organism anatomy and cell type	Cell ontology	CL	A controlled vocabulary of cell types. Not organism-specific
Organism anatomy and cell type	Foundational Model of Anatomy	FMA	A reduced version of the FMA with only the *is_a*, *part_of* and *has_part* relations included.
Phenotypic qualities	Phenotype ontology	PATO	Ontology of phenotypic qualities/properties intended for use with other ontologies such as organism-specific anatomies.
Biochemical	Chemicals of biological interest	CheBI	Classification of the chemicals with relevance in the biological domain. Can be used in conjunction with other ontologies.
Experimental measurement	Ontology for Biomedical Investigation	OBI	Ontology for describing the design of an experiment, the protocols, materials and instrumentation used, the data generated and the type of analyses performed
Experimental measurement	Unit	UNIT	Metric unit ontology intended for use with PATO
Relations	Relationship ontology	RO	Ontology of the relations being used by the OBO ontologies

15.2.2 The Gene Ontology

The Gene Ontology (Ashburner et al., 2000) provides three structured, controlled, non-organism specific vocabularies describing the entities that exist in the domains of molecular function, cellular location and biological processes of genes or gene products.

The project began in 1998 as a collaboration between the curators of three of the major model organism databases (FlyBase, the Mouse Genome Informatics database, and the Saccharomyces Genome Database), and arose out of the need for these communities to share a common, unambiguous vocabulary for functional annotation of genes and gene products within these databases.

The aims of the Gene Ontology consortium, which has since expanded to include 16 members, are: (i) to develop a set of controlled, structured vocabularies to describe key domains of molecular biology, including gene product attributes and biological sequences; (ii) to apply GO terms in the annotation of sequences, genes or gene products in biological databases; and (iii) to provide a centralized public resource allowing universal access to the ontologies, annotation data sets and software tools developed for use with GO data. (Harris et al., 2004). The success of the GO is evidenced by its widespread adoption. Using the search term "Gene Ontology" identifies more than 1,843 citations in GoogleScholar in June 2007 It was the success of the Gene Ontology that inspired the development of a large number of domain ontologies, many of which are now gathered under the umbrella of the OBO consortium. In understanding the reasons for this success it is important to note that the GO consortium focused on openness and community-involvement, and the application to real data as key principles in the development, and that these, together with others factors discussed in an opinion article (Lewis, 2005) have proven extremely powerful motivators for the biomedical community.

The entities captured in each of the three ontologies that compose the Gene Ontology have *is-a* and *part-of* relations to other entities (Fig. 15.1). There is no explicit link between the three ontologies that make up the Gene Ontology, although relationships between the three ontologies exist. There have been various approaches to making these relationships explicit (Bodenreider et al., 2003; Bada and Hunter, 2007). The three ontologies are generally represented as directed acyclic graphs, so that multiple inheritance is possible. A large number of the other OBO ontologies are also represented as DAGs.

15.2.3 Ontology Representation

Many biomedical ontologies are made available in the OBO flatfile format (Golbreich and Horrocks, 2007). The OBO flatfile format in its original form specifies a directed acyclic graph (Fig. 15.2). In this graph, labeled nodes represent categories, labeled edges relationships between categories.

The Gene Ontology was the first ontology to use this representation format, together with one semantic rule, the True Path Rule. The true path rule states that

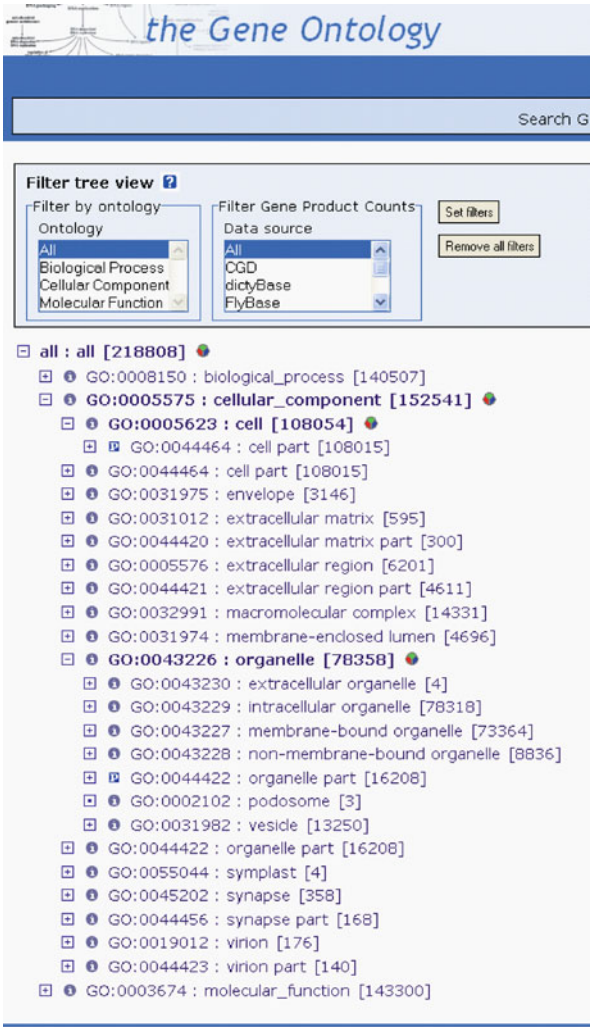

Fig. 15.1 A screenshot of a subsection of the Gene Ontology using the AmiGO browser.[5] Relationship types are indicated to the *left* of the term accession as (I) for *is-a* or (P) for *part-of*. The *cellular compartment* ontology is expanded to show entities and their relationships. An additional feature is the ability to view genes and gene products which have been annotated with each term. The number in square brackets following the term name indicates how many gene products in public databases have been annotated using the term

"the pathway from a child term all the way up to its top-level parent(s) must always be true" (Ashburner et al., 2000). In the beginning, this rule was applied to the annotations of categories in the Gene Ontology: an annotation to a category remained

[5] http://amigo.geneontology.org/

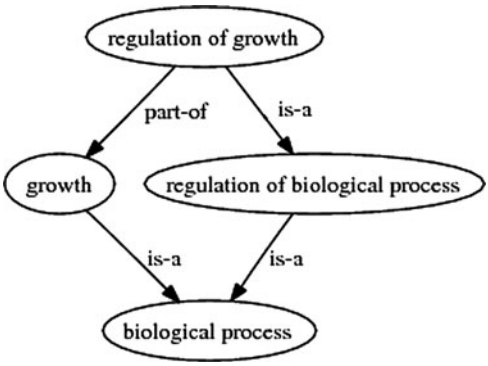

Fig. 15.2 A part of the graph structure of the gene ontology. Nodes represent categories and edges are relations between the categories. The figure shows four categories, linked using *is-a* and *part-of* relations

a valid annotation for all *is-a* and *part-of* parents of the category. The annotation relation is not an ontological relationship and may have varying meanings. Therefore, a more precise definition and semantics for these DAGs was developed in first order logic (Smith et al., 2005) and description logics (Golbreich and Horrocks, 2007).

Let C be a set of concept names, $R \supseteq \{is-a\}$ a set of relationship names, G = (V,E,c,r) be a labeled graph with vertices V, edges $E \subseteq V \times V$, a function $c:V \rightarrow C$ and a function $r:E \rightarrow R$. Then, G is equivalent to a theory T in first order logic over the signature $\Sigma = (\{:\} \cup R \cup C)$ such that for each $e \in E$:

(1) If r(e)=is-a and e=(a,b) with c(a)=c1 and c(b)=c2, then
$\{\forall x\,(x:c1 \rightarrow x:c2)\} \in T$
(2) If r(e)=S and e=(a,b) with c(a)=c1 and c(b)=c2, then
$\{\forall x\,(x:c1 \rightarrow \exists y\,(y:c2 \wedge S\,(x,y)))\} \in T$

The relationship ":" denotes the binary instantiation relation between an individual and a category. S denotes the additional relations used in this ontology. For example, if R={*part-of*, *is-a*} as in the Gene Ontology, an edge e=(c1,c2) with the label r(e)=*part-of* is translated to: forall x (x:c1 -> exists y (y:c2 AND *part-of*(x,y))).

One consequence of this definition and the True-Path Rule is that the *part-of* relation is transitive. Another important consequence is that the relation represented in the DAG is a necessary relation: there are no exceptions. The translation of a DAG into first order logic was not known from the beginning. Many of the criticism of the Gene Ontology and similar ontologies arose from misunderstandings of relations between categories. For a detailed discussion, see Sections 15.3, 15.4.1, and 15.4.4.

Recently, Semantic Web Technology is used for the development of biomedical ontologies. In particular, most ontologies that have commonly be represented in the OBO format as DAG are now available in OWL. Newly developed ontologies are often developed using a more expressive knowledge representation format, such as OWL.

15.2.4 Ontology Curation

Development and curation of the bio-ontologies is generally performed by domain experts, and consultation with ontologists is becoming more frequent. The OBO Foundry, for example, welcomes community input into OBO ontologies, and suggestions for changes or additions are implemented after careful evaluation by the curators. This process ensures that the ontology is a stable, versioned resource of high quality and consistency. Alternative models for ontology curation, including direct community curation via a wiki interface have been proposed (Hoehndorf et al., 2006) but are not yet widely adopted, largely due to concerns over the decrease in quality and increase in inconsistencies that may result if curation was completely unrestricted.

15.2.5 Annotation

A distinguishing feature of many biomedical ontologies is that they have been developed for specific use in the annotation of biomedical data such that this data can be shared and integrated. Annotation is the process whereby the terms from an ontology are associated with some experimental data (Fig. 15.3). For example, terms from the Gene Ontology have been used to describe the function, cellular location and biological process involvement of the genes and gene products in multiple model organism databases.

Annotations are contributed by consortium members and independent researchers. In the Gene Ontology the annotation data is generated largely by the collaborating model organism databases which then contribute these annotations to GO for storage and distribution. Each GO annotation has metadata identifying (i) who made the association between gene and GO term, (ii) the evidence supporting the association, and (iii) when the association was made.

Each association is labeled with an "evidence code" indicating the type of evidence that supports that association being made. Distinguishing between types of support for an association allows researchers using the data to decide how much confidence to place in the annotation. A large number of the annotations in the GO database are extracted from the biomedical literature by curators who read and interpret the statements about gene function and localization that are made in scientific papers. While manual curation provides the highest quality associations, it is time-consuming and dependent on skilled biologists. As a result high-throughput methods to associate annotations with genes/gene products using electronic methods have been developed. These approaches include extraction of associations from literature using text mining approaches, or the transfer of annotation from genes known to have similarity in their DNA sequence or protein structure. Direct experimental evidence confirmed by a human curator is generally considered more convincing than inference from automated analyses or associations based on sequence or structural similarities which have not been reviewed by a curator.

"Using microarray analysis, we identified RERG (ras-related and estrogen-regulated growth inhibitor). Like Ras, RERG protein exhibited intrinsic GDP/GTP binding and GTP hydrolysis activity. Unlike Ras proteins, RERG lacks a known recognition site for COOH-terminal prenylation and was localized primarily in the cytoplasm."

Text string	GO Ontology	GO Term	GO ID
estrogen-regulated	process	response to hormone stimulus	GO:0009725
growth inhibitor	process	negative regulation of cell growth	GO:0030308
Ras, GDP/GTP binding	process	small GTPase mediated signal transduction	GO:0007624
GDP/GTP binding	function	GDP binding	GO:0019003
GDP/GTP binding	function	GTP binding	GO:0005525
GTP hydrolysis	function	GTPase activity	GO:0003924
cytoplasm	component	cytoplasm	GO:0005737

Fig. 15.3 An example of the process of annotating the protein RERG with terms from the gene ontology. Associations are made between the text of a scientific paper (*top*) and terms from the Gene Ontology biological process ontology (response to hormone stimulus, growth negative regulation of cell growth, small GTPase mediated signal transduction), molecular function ontology (GDP binding , GTP binding) and cellular component ontology (cytoplasm)

The genes/gene products from more than 35 distinct genomes have been annotated using the Gene Ontology. Additionally, the Gene Ontology Annotation (GOA) project[6] (Camon et al., 2004) provides high quality GO-based annotations of the proteins in the UniProt knowledgebase. GOA provides annotated entries for over 60,000 species, making it the largest contributor the GO annotation effort. The annotations are generated through a combination of electronic and manual techniques. A list of all the available annotations can be retrieved from the GO project website.[7] The various tools used to build the annotations are also distributed via the project website.[8]

A large number of the applications for which ontologies are used in biomedicine make extensive use of the annotations. These applications are discussed in more detail in Section 15.5 of this chapter.

[6] http://www.ebi.ac.uk/GOA

[7] http://www.geneontology.org/GO.current.annotations.shtml

[8] http://www.geneontology.org/GO.tools.shtml#annot

15.3 Criticism and Extension of the Gene Ontology

The Gene Ontology was criticized in a series of articles (Kumar et al., 2003; Smith et al., 2003; Kumar and Smith, 2004; Smith et al., 2004). Major confusion arose from the fact that, despite its name, the Gene Ontology is viewed by its curators as a controlled vocabulary rather than as an ontology. Important features of an ontology are missing from the Gene Ontology, most notably a formal specification and definition of the categories and relations in a formal language like description logic or first order logic. Although ontological notions such as *part*, *function*, *process*, and *object* are used in the names and textual definitions of the Gene Ontology's terms, none of these are properly defined.

The Gene Ontology has been further criticized for its lack of logical and ontological rigor. The representation as a directed acyclic graph was not formalized in the early stages of the project and the *part-of* relation was used in different, inconsistent ways within the ontologies. For example, organism-specific *part-of* statements were included in the Gene Ontology so that *part-of* statements were not always true, but only within the context of certain organisms.

These problems arose from two main areas. The first is a misunderstanding of the *part-of* relationship between categories. The definition of the *part-of* relationship was only added after a major part of the Gene Ontology had already been developed, with the result that *part-of* was not uniformly used according to the definition. In particular, default knowledge was included: *part-of* relations between categories which usually, but not always, hold true. The second reason for misunderstanding of *part-of* relations was the use of the same names for similar, but biologically unrelated types of entities in different organisms. An example is the fruiting body development in fungi and bacteria. In bacteria, the fruiting body development is a kind of cell communication, in fungi it is a kind of organ development. Whenever similar terms are used for different phenomena in different organisms, the Gene Ontology makes these terms organism specific by adding a "sensu" statement to the term. Therefore, fruiting body development (sensu Fungi) and fruiting body development (sensu Bacteria) are two different categories within the Gene Ontology.

Further analysis of the Gene Ontology revealed the implicit ontological distinctions made in its three disjoint taxonomic trees. There are three disjoint taxonomies: Cellular Component, Biological Process and Molecular Function. While cellular components are identified as subclasses of "substance", the distinction between Biological Process and Molecular Function proved to be more difficult. In particular, the Biological Process taxonomy contained terms such as "transport", while the Molecular Function taxonomy contained "transporter". In 2003, a major renaming of the terms in the Molecular Function taxonomy occurred, adding "activity" to the end of each term to reflect more closely the dynamic character of the terms described. However, the relationship between the Molecular Function taxonomy and the Biological Process taxonomy remained unclear, as did the exact nature of terms described in the Molecular Function taxonomy. The definition that relates Molecular Function to Biological Process is that a biological process is series of events accomplished by one or more ordered assemblies of molecular functions. This suggests

that the activities described in the Molecular Function taxonomy are a part of some biological process.

The analysis of the Gene Ontology according to the top-level ontological distinctions of the Basic Formal Ontology (BFO) (Grenon et al., 2004) concluded that cellular components are a subclass of the BFO's continuant hierarchy, while biological processes and molecular functions as they are defined by GO are occurrents. It also concluded that molecular functions in GO are not a subclass of the function category in BFO, which are dependent continuants, but rather functionings.

A further analysis of the relationship between functions and processes in the Gene Ontology was performed using the Ontology of Functions (OF), a top-level ontology of functions (Burek et al., 2006). The OF provides a framework for defining the structure of functions, the function's relation to processes and to objects

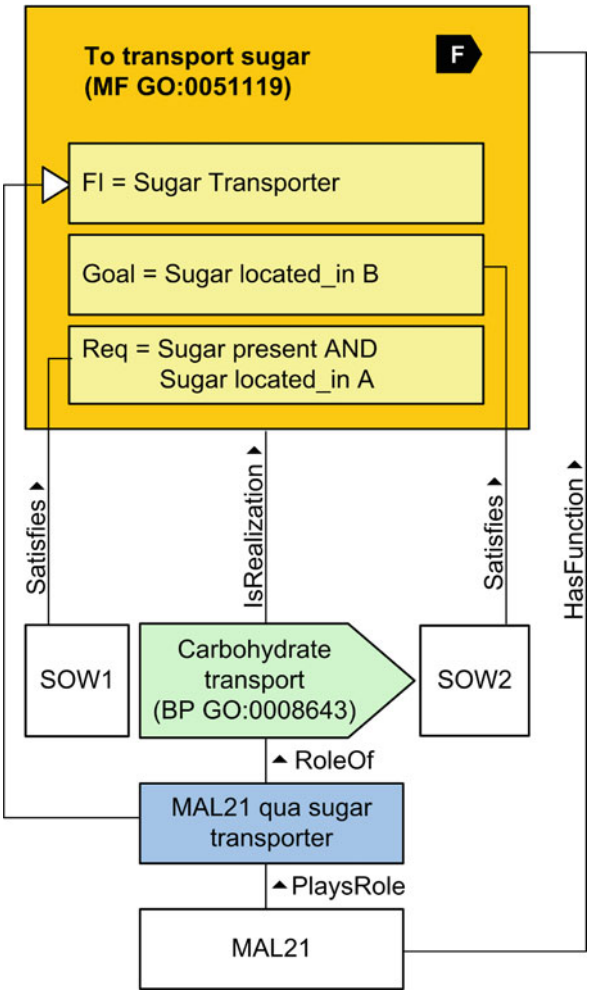

Fig. 15.4 The function "to transport sugar" represented in the framework of the ontology of functions. The function is represented using requirements and a goal. The functional item "sugar transporter" is the role that the function bearer (MAL21) plays in any realization of the function. The process "carbohydrate transport" (from gene ontology's process classification) is the realization of the function, entities of the type MAL21 are bearers of the function

that have a function ascribed to them. According to the analysis performed in OF and illustrated in Fig. 15.4, functions are defined by means of requirements, goals and a functional item. Requirements correspond to initial conditions which must be satisfied whenever a function is realized. The goal is supposed to be achieved by the function. The functional item is a role (Loebe, 2005) played by some entity in any realization of the function. For example, in the function "to transport sugar", illustrated in Fig. 15.4, the functional item is a sugar transporter role. The realization of a function is an entity which provides a transition from the state of the world in which the requirements of the function are fulfilled, to the state in which the goal of the function is fulfilled. This will usually be a process such as "sugar transport", but may be any other entity. The functional item must be played in the realization of the function. The entity playing this role in the realization is the function bearer. Applied to the GO, this yields a complete picture covering all of the GO's taxonomies, and its annotated data. The molecular function taxonomy describes the functions of gene products. These functions are realized by categories taken from the GO's biological process taxonomy. Cellular components may participate in these processes, potentially bearing a function. However, most of the molecular functions covered by the GO are functions of the gene products that are annotated to the function category. Gene products are the bearers of the functions, and they play the role of the functional item in the realization of the function.

15.4 Biomedical Ontology Integration Through the Application of Ontological Design Principles

With the increasing number of biomedical domain ontologies there is a need for a common ontological framework in which these ontologies can be integrated. The majority of ontologies that are currently available have been developed separately, and while many adhere to the OBO guidelines this has not yet guaranteed that they are fully interoperable. There are therefore several independent efforts that attempt to integrate multiple biomedical domain ontologies.

Two different approaches are taken towards the integration of ontologies in biomedicine. The first attempts to construct upper domain ontologies based on a top-level ontology. The second constructs a core ontology with which the domain ontologies are then aligned.

Upper domain ontologies define the most general categories within a domain using the categories of a top-level ontology. For example, the category "Material structure" may be specialized to "Cell" or "Molecule", imposing additional restrictions on these categories. A core ontology attempts to define the scope of a domain. In particular, it identifies the "core"; concepts of a domain and specifies the relation of each category or sub-domain to this "core".

We discuss three ontologies that can be used to integrate biomedical ontologies: the BioTop (Schulz et al., 2006) ontology together with the OBO Relationship

Ontology (Smith et al., 2005), the Simple Bio Upper Ontology (SBUO),[9] and the General Formal Ontology-Biology (GFO-Bio).[10] The first two are upper domain ontologies, the latter is both an upper domain ontology and a core ontology.

15.4.1 The OBO Relationship Ontology

The OBO Relationship Ontology (OBO-RO) (Smith et al., 2005) is an ontology of the relationships that are used between entities in biomedical ontologies. Its basic ontology contains only two categories, Continuant and Occurrent. Continuants are entities which are wholly present at a single point in time, while occurrents have temporal parts and unfold through time. The OBO-RO provides a set of basic relations and gives axioms for these. Among the relations provided in the OBO-RO are the *is-a* relation, various mereotopological relations, participation, and transformation and derivation relations. For each relation, axioms specifying reflexivity, transitivity and symmetry are provided. In addition, further definitions are given in English text.

Because the OBO Relationship Ontology attempts to provide a unifying framework for all biomedical ontologies, the axioms for the relations are weak compared to more specialized theories. For example, the axioms for the part-of relationship are reflexivity, transitivity and anti-symmetry.

A number of relations are defined which are intended for use only within the biomedical domain. Among them are the relation *transformation_of* and *derives_from*. The *transformation_of* relation is a relation between two identical biological individuals at two different points in time. The *derives_from* relation relates two distinct individuals at two different points in time, and the later individual is a result of either division or fusion of the previous individual.

The OBO-RO was developed at a time when most biomedical ontologies were available as directed acyclic graphs. In these graphs, relations such as *part-of* were used as inconsistently and ambiguously. By providing these relations with consistent and unambiguous definitions, the OBO-RO aims to facilitate ontology inter-operability and to support advanced reasoning across these ontologies. New ontologies in the OBO library are required to comply with the OBO-RO.

15.4.2 BioTop and the Simple Bio Upper Ontology

The BioTop Ontology (Schulz et al., 2006) is a further development of the GENIA upper ontology. GENIA is an ontology that is intended for use in semantic annotation of texts in biological text mining (Kim et al., 2003). Several problems with GENIA's upper ontology have been identified. BioTop is an upper domain ontology for biology based on the top-level ontology BFO (Grenon, 2003), with

[9] http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio/
[10] http://onto.eva.mpg.de/gfo-bio.html

some concepts borrowed from DOLCE (Masolo, Borgo et al.). The relationships used in BioTop are the ones used in the OBO Relationship Ontology, plus some additional relations.

Like GENIA's upper ontology, BioTop is mainly an ontology of continuants: entities that are wholly present at each point in time at which they exist, and preserve their identity through time. Axioms are given in OWL-DL for the upper categories used in biomedical domain ontologies. For example, the category *Cell* is defined as having some Cytoplasm and no Cell as part, and having some CellularComponent and some Membrane as component.

BioTop is intended to be applied as an upper level ontology for all the ontologies listed under the OBO umbrella. By providing definitions for the upper categories of these ontologies, BioTop enforces ontological rigor and attempts to eliminate ambiguities in the use of categories. For example, when two ontologies include a Cell category, and both use BioTop for defining this Cell category, interoperability between these ontologies is made simpler.

The Simple Bio Upper Ontology (SBUO) is an upper domain ontology like BioTop. It is mainly founded in the DOLCE top-level ontology, with some ideas from BFO included. Due to the top-level ontology used, several differences distinguish the two ontologies. In particular, biological sequences like DNA sequences are abstract individuals in SBUO, while they are modeled as subclasses of molecules in BFO.

15.4.3 GFO-Bio

While BioTop and SBUO are upper domain ontologies, GFO-Bio[11] is both an upper domain ontology and a core ontology. This is a result of the fact that GFO-Bio attempts to make the nature of the biological domain precise, and analyzes the categories used in the upper domain ontology part with respect to their relation to biology.

GFO-Bio is based on the top-level ontology GFO (Herre, Heller et al.). The relevant features of GFO-Bio's top-level ontology that allow it to be used to analyze the nature of a domain are the inclusion of a theory of levels of reality, and explicit support for higher-order categories in GFO.

In GFO, a level of reality is a higher-order category which has as instances the categories belonging to a level (see chapter on GFO and levels). The biological level is defined using the notion of "autopoiesis" as the property of living systems (Maturana and Varela, 1991). Using the concept of autopoiesis, two principal categories are identified: Cell and Organism, which both exhibit the property of autopoietic systems on the material stratum. To the instances of these principal categories, relations taken from the top-level ontology GFO are applied to yield further categories. For example, an analysis of organisms using the subsumption relation results in a species tree. To each category in this tree, mereological relations may

[11] http://onto.eva.mpg.de/gfo-bio.html

be applied to obtain a classification of anatomical parts of organisms of one species. GFO-Bio structures categories according to the relationships that must be applied to its principal categories, Organism and Cell, in order to obtain the category. This approach has been called “facet analysis” in the spirit of faceted classification used in library science.

A core ontology, such as GFO-Bio, is used differently for the integration of ontologies than an upper domain ontology. While upper domain ontologies define the upper level concepts of a domain ontology using restrictions on categories, and thereby provide definitions and restrictions for the domain categories, a core ontology specifies the relation of a domain ontology to the principal categories of the core ontology. It therefore has two main purposes: to structure sub-domains within biology according to ontological principles, and to make the nature of the biological domain precise, thereby delimiting it and allowing for a structured integration within a wider ontological framework covering multiple domains, such as chemistry. Further, it allows a faceted view of the domain of biology, starting from the principle categories of biology and exploring different facets – relationships – of these principle categories. A part of the taxonomic tree of GFO-Bio depicting the biological level and several facets or sub domains within this level is shown in Fig. 15.5. The material stratum is a higher-order category, and the biological level a sub-category of the material stratum.

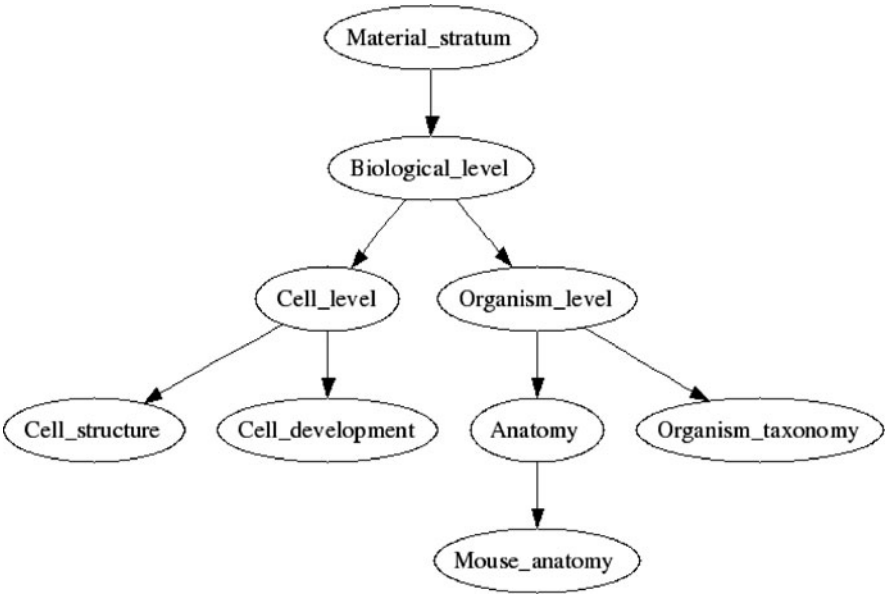

Fig. 15.5 A fragment of GFO-Bio’s classification of the biological level. The Material_stratum is considered a subcategory of GFO’s “Category”. Instances of the Material_stratum category are the categories that belong to the material stratum. Similarly, the Biological_level category has the categories and relations pertaining to the biological level as instances. The biological level is further refined to more detailed sub-levels or domains such as Anatomy

15.4.4 Defaults and Exceptions for Ontology Interoperability

Some biomedical ontologies take a particular view on the domain they cover in that they describe idealizations of the domain. Most of the anatomical ontologies fall into this class. For example the Foundational Model of Anatomy (Rosse and Mejino, 2003) describes an idealized, canonical human anatomy. A separate class of ontologies describes phenomena within a domain where these phenomena may be exceptions. An example is the Mammalian Phenotype Ontology, which is specifically designed to describe abnormal mouse phenotypes which arise from genetic modifications in mice.

Interoperability between these types of ontologies would facilitate the consistent use of biomedical data in the form of annotations, allow for queries over multiple ontologies and form a rich knowledge resource for biomedicine that could be further used in solving problems and stating hypotheses. The absence of clear principles for achieving interoperability between ontologies of this kind hinders the development of advanced applications and analysis tools based on these ontologies. As we will show in the following section, interoperability of ontologies of these different types cannot be achieved by the methods developed hitherto, and a new set of methods that transcends the framework of classical logic must be introduced in order to avoid inconsistencies and at the same time preserve the specificity of both types of knowledge.

A canonical anatomy ontology such as the Foundational Model of Anatomy contains rules such as every instance of a human body has as part an appendix. (1)

This rule does not necessarily apply to every real human body: an individual human body may *lack* an appendix as part. However, the rule describes an idealized or *canonical* human. Phenotype ontologies may describe exceptions to these idealizations. For example, an individual may both be an instance of a human body as described in the FMA (which implies an appendix as part) and an instance of the category "human body with absent appendix". In a classical logical framework, such as those used in the OBO Relationship Ontology or in the form of the Web Ontology Language (OWL) (McGuinness, 2004), a formalization of these two statements would lead to an inconsistency. A human body in the former case has an appendix as a part, while in the latter case it does not. Instantiating both by an individual causes the inconsistency. A logical inconsistency in the formal sense can only arise when the logical functor of negation is used. This functor is hidden in concepts such as "absent X", as used in the Mammalian Phenotype Ontology (Smith, 2005).

In order to avoid terms such as "absent X" and make the negation explicit, the **lacks** relation was introduced (Ceusters, 2007), which can be explicitly defined as:

Individual *p* **lacks** category *C* with respect to relation **R** if and only if there is not an *x* such that: *x***R***p* and *x* is an instance of *C*.

It is possible to use binary relations of the kind *x* **lacks-R** *C* instead of *x* **lacks** *C* with respect to **R**. For example, the fact that some individual *x* **lacks** a category *C* with respect to the relation **has-part** will be denoted as *x* **lacks-part** *C*.

The use of the **lacks** relation may cause an inconsistency when a canonical ontology and a corresponding phenotype ontology are used together with instances in a

classical logic formalism, such as first order logic or description logic. The reason is that classical formalisms enforce very strict interpretations, e.g. of quantifications like "every human", which results in *monotonicity* of these formalisms: the inferences drawn from a classical logical theory *T* remain true in every extension of *T* with additional facts. In order to prevent inconsistencies, while at the same time preserving the intuition behind statements such as "a human has an appendix as part", such statements in the canonical ontology must be weakened. What is required is a *nonmonotonic* logic with which the statements in a canonical ontology can be treated as true by default, but adding additional knowledge, by reference to a phenotype ontology or using a statement involving the **lacks** relation (and therefore negation), may invalidate the conclusions previously drawn.

In order to describe the nature of default relationships between two categories, new relations must be introduced, such as **CC-canonical-has-part**. For each relationship **R** between individuals, a set of relations is introduced according to Table 15.2. Then, the relationship between "human" and "appendix" becomes "human **CC-canonical-has-part** appendix". Further, this relationship corresponds to a *default rule*:
forall x, C1, C2: if C1 CC-canonical-has-part C2 and x IC-instance-of C1 then

by default: there exists a y: y IC-instance-of C2 and x II-has-part y.

Defaults rules can be formalized using answer set programs. Answer set programs are logic programs that employ two kinds of negation, strong and weak. While strong negation corresponds to classical negation, weak negation is also referred to as "default negation". Intuitively, the weakly negated statement "not A" means "A cannot be proven".

Answer set programs must be further combined with ontology representation, in order to be used within ontologies. For example, the system DLVHEX allows for a bidirectional flow of information between an answer-set program and a description logic knowledge base (Eiter, 2006). Relationships that are used in an ontology are made available to the DLVHEX system. Then, it is possible to express the necessary

Table 15.2 A schema of the relations introduced. Domain and range for the relations are encoded in the prefix of their name (e.g., IC means that the domain is Individual and the range Category). For each relation that is used in an imported ontology, a number of relations between categories, individuals, and between individuals and categories can be created. The CC-canonical-R relationship is a default relation which is accompanied by axioms in an answer set program in order to describe its semantic as a default

Relationship	Definition
x **II-R** *y*	Individuals *x* and *y* stand in the primitive relation **II-R**.
x **IC-R** *y*	There exists an individuals *z*, which is an instance of *x*, such that *x* **II-R** *z*.
x **CC-R** *y*	For all individuals *a* which are an instance of *x*: *a* **IC-R** *y*.
x **CC-canonical-R** *y*	For all individuals *a* which are an instance of *x*: normally, *a* **IC-R** *y*.
x **II-lacks-R** *y*	Not *x* **II-R** *y*.
x **IC-lacks-R** *y*	Not *x* **IC-R** *y*.
x **CC-lacks-R** *y*	For all individuals a such that: *a* **IC-instance-of** *x*, *a* **IC-lacks-R** *y*.

axioms for relations of the kind **CC-canonical-R**. For example, for the relationship **CC-canonical-has-part**, the following axiom can be added:

```
IC-has-part(X,Y) :- ind(X), class(Y), inst(X,Z),
                    CC-canonical-has-part(Z,Y),
                    not IC-lacks-has-part(X,Y), class(Z).
```

This means that if two categories *Z* and *Y* stand in the relation **CC-canonical-has-part**, and *it is not provable that X* **IC-lacks-has-part** *Y* (`not IC-lacksHasPart(X,Y)`), then it is concluded that an individual *X*, which is an instance of *Z*, stands in the relation **IC-has-part** to the category *Y*.

Extending biomedical ontologies with the capability for non-monotonic reasoning allows for interoperability between ontologies describing canonical knowledge within a domain and phenotype ontologies (which describe phenomena). Using a hybrid approach by combining traditional ontology representation languages such as OWL or OBO DAGs with answer set programs allows for the reuse of tools that are used in ontology development, such as Protege (Noy et al., 2003) or OBO-Edit (Day-Richter et al., 2007).

15.5 Applications

Development of the majority of the bio-ontologies has been driven by the need to order and analyze the vast amount of data collected in biological databases and acquired by experiments. The biological community has actively applied ontologies for the annotation of biological data types. A feature that distinguishes the biomedical ontologies is the vast amount of experimental data that is annotated using these ontologies. It is the combination of the ontologies with this data that has enabled large-scale biological description and discovery. A number of software packages supporting a variety of biological applications have been developed by the community, only a few of which we will discuss here. A software repository for of these packages is maintained by the Gene Ontology Consortium,[12] and while some of the tools are GO-specific, some can be used with multiple bio-ontologies.

15.5.1 Annotation and Retrieval of Data

Through formalizing the terms used for a domain and then using these for the annotation of biological data such as genes and proteins, the bio-ontologies have provided researchers with the ability to browse and retrieve data according to well known terms. In an initial approach to help researchers in genetics manage the rising number of sequences in public databases, controlled vocabularies were used

[12] http://www.geneontology.org/GO.tools.shtml

to assign commonly used terms to genes and proteins (eg:. The protein database Swiss-Prot (Boeckmann et al., 2003)) or cDNA libraries (Kelso et al., 2003). These controlled vocabularies were later supplemented by bio-ontologies to provide researchers with domain-specific hierarchies for the browsing and retrieval of data. For this approach no more than a simple is-a hierarchy is needed, giving a possible explanation for the simple structure of the OBO Ontologies. A standard example of a pure is-a hierarchy in biology is the classification of species, constituting an integral part in the organization of genetic information (Wheeler et al., 2003). Biological databases now make extensively use of bio-ontologies to provide controlled terms for the description of various aspects of genes and proteins.

15.5.2 Statistical Analysis of Experiments

Current technologies allow for massive parallel measurements in genetic experiments. One well-known and widely used form of experiment measures the relative amount of transcript from DNA for several thousand genes on a microarray chip (reviewed in Lockhart and Winzeler, 2000). However, the analysis and interpretation of the data generated by these experiments is often hampered by two major problems:

1. the power to draw a significant conclusion from a single measurement is low because of large technical variance in the experimental measurements, and
2. data on the level of single genes does not allow for a direct insight into the affected higher level functions of the organism.

These problems led to the development of several applications (eg:. GOstat (Beissbarth and Speed, 2004) or FUNC[13] (Pruefer et al., 2007) which make use of the simple DAG structure of the Gene Ontology in order to group genes by their annotation. This grouping increases the power to detect differences, as the measurements for multiple genes can be combined for testing. Additionally, the statistical test on the Gene Ontology DAG results in a list of significant groups. These groups are described by meaningful terms from the Gene Ontology, thus helping the user to interpret the result in terms of the biologically relevant affected processes and functions as well as the cellular localization.

While these applications vary in respect to the implemented statistical tests and the user interface, their general method is very similar. As a prerequisite, genes need to have an assigned value as the result of the experiment. These values are then collected in Gene Ontology groups according to the annotation of the gene and can be propagated to linked higher level groups in the DAG because of the True Path Rule. An appropriate statistical test is then applied to each group. Since many groups are tested for significance, the chance of a false positive result is not at the desired probability of error. This constitutes a well known problem of statistics known as multiple hypothesis testing and is addressed in several of these packages using

[13] http://func.eva.mpg.de/

a variety of methods for correction (Manly, 2004). The approach that we have described here is not limited to the Gene Ontology, but can be applied to any ontology that can be represented as a DAG.

15.5.3 Automatic Annotation and Community-Developed Ontologies

Given the amount of experimental and computational research required to describe gene function, the genetic bases of complex diseases, or the evolutionary history of organisms, genetics tends to be a field with a vast number of publications, often in highly focused research areas. Since the curation model used by most of the bio-ontologies requires curators to read literature in order to extract the ontological terms and annotations, this leads to a bottleneck in the curation of these ontologies. Generally the curators read a defined subset of publications to create annotations. Two alternative approaches to addressing this challenge have been undertaken.

15.5.3.1 Automatic Annotation

Using methods from computer linguistics and information extraction several authors (Hirschman, 2005) have explored automating the search for relevant publications for each term in an ontology. Information extraction from biological texts is a powerful means to increase the coverage of ontologies and their annotations. Such approaches may also have the ability to verify their correctness, providing increased confidence in the automatically generated results. Several sophisticated software implementations have been developed to extract information about e.g. gene and protein functions from biomedical literature (Camon, 2005). However, while information extraction from biomedical texts can quickly provide huge amounts of structured information that potentially can be added to ontologies as categories, relationships or annotations, manual verification and quality assurance based on human input is always beneficial.

15.5.3.2 Community Development

A very recent development to increase the amount of captured information from publications is the use of Wikis (Leuf and Cunningham, 2001) which aim to involve the community directly in the curation process. While no fully fledged Wiki for this purpose exists currently, there are several proposed methods, spanning several degrees of formalism for the captured information. A natural way of applying the current Wiki technology is to allow natural language descriptions for each gene, to supplement the genome databases with further information gained from experts (Wang, 2006). Such a wiki does not yet exist, but there are proposals to provide such functionality via a new project called WikiProteins (Giles, 2007).

The most formal approach to date is currently under development by Hoehndorf et al.(Hoehndorf et al., 2006). Within this wiki users are able to edit annotations and add or modify concepts in the ontology. Additional to the natural-language

aspect, the wiki provides a way to add formal n-ary relations with subject, object and additional mandatory and obligatory roles.

A background Core Ontology (GFO-BIO) together with a reasoner are used to ensure than the information in the wiki is reasonably accurate. The formal entries must be typed on the basis of the Core Ontology and the reasoner is applied to limit the entries to those that are consistent with already entered information.

15.5.4 Reasoning for Experimental Hypothesis Testing

There are few advanced applications of the bio-ontologies, perhaps because many still lack the required formality to support such applications. A recent and interesting example of the use of bio-ontologies in the formulation and testing of experimental hypotheses is the Robot Scientist project (Soldatova et al., 2006). The Robot Scientist is a robotic laboratory system able to design, perform and evaluate biological experiments in a microbiological laboratory. Based on a general ontology of experiments, EXPO, (Soldatova and King, 2006) in which data and metadata about all aspects of the experiment are captured, the robot is iteratively able to formulate hypotheses, physically carry out the experiments, and then evaluate results in order to use the information gathered in the next experiment.

A second example is Hybrow (Racunas et al., 2004). HyBrow is a system to design and evaluate hypotheses and verify their consistency with available biomedical knowledge. It uses an event-based ontology for representing biological processes in the background. A prototypical implementation is available.[14]

15.6 Summary and Conclusions

The research field of bio-ontologies has grown rapidly in the past 10 years. This is a direct result of the need in the biomedical research community to define and share the vocabulary used for the description of the growing quantities of biological data being generated. With increasing amount of data more difficulties were encountered in managing, sharing and integrating these data. While several early projects, notably BioCyc (Karp et al., 2005) and GALEN (Rector and Nowlan, 1994), provided ontologies for parts of the biomedical domain, the newer, "lightweight" ontologies such as the Gene Ontology were developed by biologists to solve the specific problems that they face in daily research activities. These ontologies were therefore designed to address a specific, restricted set of problems – mainly annotation and database integration – and initially tended to sacrifice formal logical and ontological rigor to achieve this goal in a reasonable time-frame. Over time, and following the success of ontologies like the Gene Ontology, biomedical

[14] www.hybrow.org/

ontologies are being gradually extended, formal foundations laid, and ontological principles applied. This is being done in an effort to facilitate interoperability between the various ontologies that were developed for distinct, but related domains. Ultimately, these improvements will enable the automatic detection and prevention of inconsistencies, and automatic extraction of implicit knowledge. The development and application of top-level ontologies, the construction of upper-domain and core ontologies, and the unification of the relationships used in the various biomedical sub-domains are all significant steps in the construction of a unified biomedical knowledge base. As an increasing amount of knowledge is formalized, the application of ontologies and other biomedical knowledge bases for the generation of biological and biomedical hypotheses, their verification, the automatic planning and evaluation of experiments and the detection of conflicting biomedical claims may become possible. The community-wide adoption of software implementations that use ontologies for the statistical analysis of experimentally generated gene lists (Beissbarth and Speed, 2004; Prufer et al., 2007), or the identification of protein functions using ontologies (Wolstencroft, 2006) indicate that ontologically-based applications are a welcome addition to the biologists data generation and analysis toolset. Biomedicine is likely to remain a largely data-driven discipline that capitalizes on the intuition, experience and intellect of the biomedical researcher. However, parts of the field are amenable to becoming knowledge-driven disciplines. It therefore seems likely that ontology-based biomedical knowledge-bases will play an increasingly important role in modern biomedicine, and act a motivating force in computational logics, Semantic Web technologies, and for foundational ontological research.

References

Ashburner, M., et al. 2000. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics* 25(1):25–29.

Bada, M., and L. Hunter. 2007. Enrichment of OBO ontologies. *Journal of Biomedical Informatics* 40(3):300–315.

Beissbarth, T., and T.P. Speed. 2004. GOstat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 20(9):1464–1465.

Bodenreider, O., et al. 2003. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: A feasibility study. *Studies in Health Technology and Informatics* 95:379–384.

Boeckmann, B., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31(1):365–370.

Burek, P., et al. 2006. A top-level ontology of functions and its application in the open biomedical ontologies. *Bioinformatics* 22(14):e66–e73.

Camon, E., et al. 2004. The gene ontology annotation (GOA) database: Sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research* 32(Database issue):D262–D266.

Camon, E.B., et al. 2005. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6(Suppl 1):17.

Ceusters, W., et al. 2007. Negative findings in electronic health records and biomedical ontologies: A realist approach. . *International Journal of Medical Informatics* 76(Suppl 3):S326–S333.

Day-Richter, J., et al. 2007. OBO-edit – An ontology editor for biologists. *Bioinformatics* 23(16):2198–2200.

Eiter, T., et al. 2006. Dlvhex: A system for integrating multiple semantics in an answer-set programming framework. Proceedings 20th Workshop on Logic Programming and Constraint Systems (WLP 06).

Field, D., et al. 2006. Meeting report: eGenomics: Cataloguing our complete genome collection II. *OMICS: A Journal of Integrative Biology* 10(2):100–104.

Giles, J. 2007. Key biology databases go wiki. *Nature* 445(7129): 691.

Golbreich, C., and I. Horrocks. 2007. The OBO to OWL mapping, go to OWL 1.1! Proceedings of OWL-ED 2007.

Grenon, P. 2003. BFO in a nutshell: A bi-categorial axiomatization of BFO and comparison with DOLCE.

Grenon, P., et al. 2004. Biodynamic ontology: Applying BFO in the biomedical domain. *Studies in Health Technology and Informatics* 102:20–38.

Harris, M.A., et al. 2004. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research* 32(Database issue):D258–D261.

Herre, H., et al. General formal ontology (GFO): A foundational ontology integrating objects and processes. Part I: Basic principles.

*Hi*rschman, L., et al. 2005. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics* 6(Suppl 1):S1.

Hoehndorf, R., et al. 2006. A proposal for a gene functions Wiki. On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, 669–678.

Karp, P.D., et al. 2005. Expansion of the bioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* 33(19): 6083–6089.

Kelso, J., et al. 2003. eVOC: A controlled vocabulary for unifying gene expression data. *Genome Research* 13(6A):1222–1230.

Kim, J.D., et al. 2003. GENIA corpus-semantically annotated corpus for bio-textmining. *Bioinformatics* 19 Suppl 1:i180–i182.

Kumar, A., and B. Smith. 2004. Enhancing GO for the sake of clinical bioinformatics. Proceedings of Bio-Ontologies Workshop.

Kumar, A., et al. 2003. The unified medical language system and the gene ontology: Some Critical Reflections. *KI2003: Advances in AI*: 135–148.

Leuf, B., and W. Cunningham. 2001. *The wiki way: Quick collaboration on the web*. Boston, MA: Addison-Wesley.

Lewis, S. E. 2005. Gene ontology: Looking backwards and forwards. *Genome Biology* 6(1):103.

Lockhart, D.J., and E.A. Winzeler. 2000. Genomics, gene expression and DNA arrays. *Nature* 405(6788):827–36.

Loebe, F. 2005. Abstract vs. social roles: A refined top-level ontological analysis. Proceedings of the 2005 AAAI fall symposium roles, An Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems', AAAI.

Manly, K.F., D. Nettleton, et al. 2004. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Research* 14(6):997–1001.

Masolo, C., et al. 2003. Wonderweb deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology.

Maturana, H.R., and F.J. Varela. 1991. *Autopoiesis and cognition: Realization of the living (Boston studies in the philosophy of science)*. Berlin: Springer.

McGuinness, D.L., and V.H., Frank. 2004. OWL web ontology language overview.

Noy, N.F., et al. 2003. Protege-2000: An open-source ontology-development and knowledge-acquisition environment. *AMIA Annual Symposium Proceedings*: 953.

Pruefer, K., et al. 2007. FUNC: A package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 8:41.

Prufer, K., et al. 2007. FUNC: A package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 8:41.

Racunas, S.A., et al. 2004. HyBrow: A prototype system for computer-aided hypothesis evaluation. *Bioinformatics* 20(Suppl 1):i257–i264.
Rector, A.L., and W.A. Nowlan. 1994. The GALEN project. *Computer Methods and Programs in Biomedicine* 45(1–2):75–78.
Rosse, C., and J.L. Mejino, Jr. 2003. A reference ontology for biomedical informatics: The foundational model of anatomy. *Journal of biomedical informatics* 36(6):478–500.
Schulz, S., et al. 2006. Towards an upper level ontology for molecular biology. *AMIA Annual Symposium Proceedings*: 694–698.
Schulze-Kremer, S. 1998. Ontologies for molecular biology. *Pacific Symposium on Biocomputing* 3:695.
Schulze-Kremer, S. 2002. Ontologies for molecular biology and bioinformatics. *In Silico Biology* 2(3):179–193.
Smith, B., et al. 2003. The ontology of the gene ontology.
Smith, B., et al. 2004. On the application of formal principles to life science data: A case study in the gene ontology. Proceedings of DILS 2004 (Data Integration in the Life Sciences), Springer.
Smith, B., et al. 2005. Relations in biomedical ontologies. *Genome Biology* 6(5):R46.
Smith, C.L., et al. 2005. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*ogy 6(1):R7.
Soldatova, L.N., and R.D. King. 2006. An ontology of scientific experiments. *Journal of the Royal Society Interface* 3(11):795–803.
Soldatova, L.N., et al. 2006. An ontology for a robot scientist. *Bioinformatics* 22(14):e464–e471.
Wang, K. 2006. Gene-function wiki would let biologists pool worldwide resources. *Nature* 439(7076):534.
Wheeler, D.L., et al. 2003. Database resources of the national center for biotechnology. *Nucleic Acids Research* 31(1):28.
Whetzel, P.L., et al. 2006. The MGED ontology: A resource for semantics-based description of microarray experiments. *Bioinformatics* 22(7):866–873.
Wolstencroft, K., et al. 2006. Protein classification using ontology classification. *Bioinformatics* 22(14):e530–e538.