Formal Ontology in Information Systems
R. Ferrario and W. Kuhn (Eds.)

doi:10.3233/978-1-61499-660-6-299

Large-Scale Reasoning over Functions in Biomedical Ontologies

Robert HOEHNDORF [a,1], Liam MENCEL [a] Georgios V GKOUTOS [b,c] and Paul N SCHOFIELD [d]

[a] *King Abdullah University of Science and Technology, Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, 4700 KAUST, Thuwal 23955-6900, Kingdom of Saudi Arabia*
[b] *College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, B15 2TT, United Kingdom*
[c] *Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, B15 2TT, United Kingdom*
[d] *Department of Physiology, Development & Neuroscience, University of Cambridge, Downing Street, CB2 3EG, United Kingdom*

Abstract. A large number of biomedical resources have been developed to represent the functions of biological entities, and these resources are widely used for data integration and analysis. Expressing functions in biomedical ontologies currently uses formal representation patterns that renders basic reasoning tasks to fall in complexity classes beyond polynomial time, thereby limiting the potential of using knowledge-based methods for data integration, querying or quality control. Here, we propose an alternative representation pattern for expressing knowledge about biological functions, together with a biological and ontological justification, which can be expressed using the description logic EL++ and implemented using the OWL 2 EL profile. To demonstrate the utility of our account of biological functions, we apply it to all proteins contained in the SwissProt database and evaluate its utility with respect to answering complex queries as well with respect to the classification and query times.

Keywords. protein, biological function, tractable reasoning, Big Ontologies

1. Introduction

The notion of biological function is widely used in the life sciences, and functions are assigned to entities ranging from whole organisms, such as worker ants, to small molecules, such as drugs or regulatory RNA. Functions are closely related to causation [1], and attributing a function to an entity provides information about *why* the entity exists as well

[1]Corresponding Author: Robert Hoehndorf, King Abdullah University of Science and Technology, Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, 4700 KAUST, PO Box 2882, Thuwal 23955-6900, Kingdom of Saudi Arabia; E-mail: robert.hoehndorf@kaust.edu.sa.

as *what* the entity can do [2]. In biology and biomedicine, large knowledge bases have been developed that contain information on the biological functions of various biological entities, such as proteins [3], regulatory RNA [4], biological pathways [5], cell types [6], or tissues and anatomical structures [7].

To formally characterize and classify functions in biology and biomedicine, the Gene Ontology (GO) [8] has been developed as an ontology of cell anatomy and biological processes. GO is arguably the most widely applied ontology in biology and biomedicine for data annotation, integration and analysis [9]. Using the GO, functions are assigned to entities (such as proteins) by asserting the *capability* of a protein to participate in particular processes – realizations of the proteins functions. Formally, asserting that some biological entities have capabilities requires the expressivity of knowledge representation languages for which basic reasoning problems, such as determining consistency or checking subsumption, cannot be decided in polynomial time [10,11,12,13]. Consequently, information about biological functions in biological knowledge bases is usually expressed using custom databases or the Resource Description Framework (RDF) [14,3] and lacks any explicit, model-theoretic semantics.

Here, we develop an ontology-based account of biological functions that is amenable to large-scale automated reasoning using the Web Ontology Language (OWL) [15], in particular using the OWL 2 EL profile [16] which supports tractable (i.e., polynomial time) reasoning. Our specific aims are to:

- create an ontologically sound representation of biological functions, specifically in the context of protein functions as described using the GO,
- identify a module of this theory that can be implemented in the description logic EL++ [17], the logic underlying the OWL 2 EL profile, and which allows for tractable automated reasoning,
- apply our framework to the UniProt knowledge base [3], a central resource in biology containing information about proteins and their functions, and
- evaluate the implementation with respect to performance and the types of queries that can be answered.

Ultimately, the resulting framework is intended to support automated consistency verification using OWL reasoners, the classification of proteins using background knowledge in ontologies (in particular the GO), and complex queries that combine information from the knowledge base with knowledge within ontologies. More importantly, it is also intended to serve as a model for how the further development of large biological knowledge bases can be improved using formal ontologies, and to enhance capabilities for quality control by adding explicit constraints whose violation can be detected automatically.

2. The Problem of Biological Functions

Biological functions have long been debated in philosophy of biology [18,2,19,20] as well as in the area of applied ontology [21,22]. At least two views of biological functions can be distinguished. First, and arguably the dominant view in the biological world, is a *causal* explanation of biological functions proposed by Wright [2] and refined by Millikan [19]. According to the causal view, an entity X has the biological function Y if

and only if X exists because it does Y [2]. This view is later refined by Millikan [19] who explicitly introduces evolutionary processes (indirectly through reproductive processes) into this account. Another view is that biological functions are *ascribed* to an entity [23,24]. In this view, entity X has the biological function Y if and only if X does Y (under some conditions, or within a certain context, C) and an agent A asserts that the purpose of X is to do Y. This view is useful in functional explanations. A more comprehensive discussion of biological functions and desiderata for ontological theories of biological functions can be found elsewhere [22].

While these theories (and others) explain how biological functions originate, we focus here on the kind of evidence that is required now, for example in the context of biological data annotation or biological literature reporting, for assigning a function to a biological entity. Here, we focus on molecular biology, in particular proteins, as these are biological entities for which new biological functions are discovered and assigned frequently. A function of a protein is established through several types of experiments that aim to establish the involvement of the protein in a biological process. For example, the function of a protein may be established by selectively removing the protein from an organism and observing the differences to an organism in which the protein has not been removed [25,26]. Importantly, since large scale evolutionary processes that have led to the development of a particular protein cannot be tested directly, a biological function is assigned based on the observation of the involvement of a protein in a process within a (repeatable) experiment. The additional assumption is that the involvement of a protein in a process rarely occurs randomly, but rather signifies a particular evolutionary history, and this justifies the assigning of a protein's "function".

3. State of the Art

In biomedical ontologies, functions are related to two kinds of entities: the function bearer C, and a process P that *realizes* a function F. The function F itself serves primarily as the reification of a – special kind [27,28] – of possibility of C to perform P. Consequently, for entities of type C with a function of type F, a common representation pattern is $\forall x(C(x) \rightarrow \exists y(F(y) \wedge hasFunction(x,y)))$ to state that all instances of C have some instance of F as function, and for functions of type F and processes P $\forall x,y : F(x) \wedge realizedBy(x,y) \rightarrow P(y)$ to state that instances of the function F are only realized by processes that are instances of P. The *realizedBy* relation is commonly taken as a primitive from an upper level ontology such as the Basic Formal Ontology [29] or the General Formal Ontology [24], and constrained by axioms or informally. For example, the function F_{gluc} to synthesize glucose from noncarbohydrate precursors (such as pyruvate, glucogenic amino acids and glycerol), which is, among others, a function of the human protein *cAMP-dependent protein kinase catalytic subunit gamma* (PRKACG), could be used in a statement such as: $\forall x(PRKACG(x) \rightarrow \exists y(F_{gluc}(y) \wedge hasFunction(x,y)))$. Similarly, the relation between the function F_{gluc} and the process P_{gluc} of *gluconeogenesis* (with identifier `GO:0006094` in the Gene Ontology) would be expressed as $\forall x,y(F_{gluc}(x) \wedge realizedBy(x,y) \rightarrow P_{gluc}(y))$.

In the OBO Relationship ontology [30], a relation (or axiom pattern [31]) *capable of* is defined between two classes, such that C is *capable of* P if and only if the following holds: $\forall x(C(x) \rightarrow \exists y(hasFunction(x,y) \wedge \forall z(realizedBy(y,z) \rightarrow P(z))))$. For example,

proteins of the type *PRKACG* are capable of facilitating *gluconeogenesis* if and only if proteins of the type *PRKACG* have a function that, if it is ever realized, is realized by processes of the type *gluconeogenesis*.

These ontology design patterns have been widely applied in biomedical ontologies when aiming to formally express the relation between structures, functions and the processes that realize them [10,11]. Most implementations are based on the Web Ontology Language (OWL) [15], and the axiom patterns fall in the *ALC* subset of OWL [32]. For example, the fact that *PRKACG* is a participant of the process *gluconeogenesis* would be expressed as the axiom $PRKACG \sqsubseteq \exists hasFunction.(\forall realizedBy.Gluconeogenesis)$. The use of both existential and universal quantification, often mixed with conjunction and disjunction, ensures that these axioms do not naturally fit in a subset (i.e., a profile) of OWL that guarantees polynomial time complexity for basic reasoning tasks such as classification or checking subsumption. Consequently, these patterns cannot easily be used for automated reasoning on a large scale, such as those found in biological and biomedical databases and ontologies, while at the same time preserving the possibility of querying efficiently the resulting knowledge using the Description Logic semantics of OWL.

To overcome these problems, modularization techniques have been developed [33]. Such techniques enable the identification of a module of an ontology, i.e., a subset of the ontology's axioms which then can be used for querying and classifying in, usually, less time than the time that would be required for querying or classifying the whole ontology. While most modularization techniques reduce the signature of an ontology, in particular locality-based modules [34], it is also possible to retain the full signature and only remove certain axioms which may contribute to reasoning complexity [13]. This second type of modularization approach can be used to reduce the complexity of querying an ontology to polynomial time while losing some inferences [13]. However, when expressing functions, or capabilities, of entities in biomedical ontologies, the type of process that would realize the function is usually the crucial and relevant aspect of the axiom that is exploited for querying and data integration; removing this axiom would no longer allow inferences about this type of process (and its properties).

In summary, there are three challenges to overcome in obtaining a tractable, large-scale representation of biological functions that can be used for automated reasoning:

- common representation patterns for biological functions in biological ontologies require an expressivity in which basic inference tasks cannot be decided in polynomial time;
- modularization techniques that reduce the signature do not enable queries of the whole ontology; and
- modularization techniques that retain the signature would remove crucial axioms required for querying knowledge about biological functions.

4. Methods

To achieve the goal of building a representation of proteins and their functions that is both ontologically sound and enables tractable automated reasoning we have two options. Either, we separately address these two tasks, starting with the ontological analysis and then applying modularization techniques to the resulting theory in order to obtain

Table 1. Overview of syntax and semantics of OWL 2 EL (omitting concrete domains). Semantics of concept descriptions and axioms is defined over a concrete domain $D = (\Delta, P)$, with Δ being a non-empty set and P a set of predicate names, and an interpretation $I = (\Delta^I, \cdot^I)$ in the standard way.

Name	DL syntax	Semantics
top	$\top$	Δ^I
bottom	$\bot$	$\emptyset$
nominal	a	$\{a^I\}$
conjunction	$C \sqcap D$	$C^I \cap D^I$
existential restriction	$\exists R.C$	$x \in \Delta^I \mid \exists y \in \Delta^I : (x,y) \in r^I \wedge y \in C^I$
general concept inclusion	$C \sqsubseteq D$	$C^I \subseteq D^I$
role inclusion	$r_1 \circ \ldots \circ r_n \sqsubseteq r$	$r_1^I \circ \ldots \circ r_n^I$

a tractable representation, or, we combine considerations related to tractability with the ontological analysis.

Both choices have advantages and disadvantages. Arguably, an ontological analysis of the domain without any constraints on tractability or particular choice of a knowledge representation language will lead to a "cleaner" and more rigorous representation of the domain since limitations of the knowledge representation language will not have to be considered. In many cases, modules can then automatically be constructed from the resulting theory so that meta-theoretical properties such as decidability or time complexity of querying the knowledge base can be obtained. On the other hand, certain ontological choices may be more amenable to modularization than others and result in more expressive knowledge bases after modularization.

Here, we chose to combine ontological analysis with considerations from knowledge representation and reasoning throughout the design of our theory of biological functions. When we are faced with different choices about how to model some phenomenon, we chose the one that is more amenable to implementation using a tractable formalism, whilst being prepared to commit to entities in our ontology that would not otherwise be required should these be deemed necessary to achieve our goals.

To formalize knowledge about biological functions, we chose the description logic EL++ [17], in the form of the OWL profile OWL 2 EL [16]. OWL 2 EL has been designed to express large ontologies while maintaining polytime reasoning for classification and instance checking [17,35], and is widely used across biomedical ontologies [36,11]. An overview of OWL 2 EL is provided in Table 1.

5. Results

Since the common ontology design pattern for expressing a capability of a biological entity falls outside of the OWL 2 EL profile, we develop an alternative representation. Using our account of biological functions above, and focusing, without loss of generality, on the case of protein functions, we informally chose the following representation: for an individual protein X to have a function of the type Y, X is required to be a member of a collection of proteins, all of the collection's members are of the same type as X, and at least one of those members actually participates in a process that realizes an instance of Y. We also extend this approach to other capabilities of proteins, in particular to the potential to be located at particular places: a protein X has the capability to be located at a location of type Y if it is a member of a collection of proteins, all members of which

are of the same type as X, of which at least one protein is actually located at an instance of Y. Figure 1 illustrates our approach.

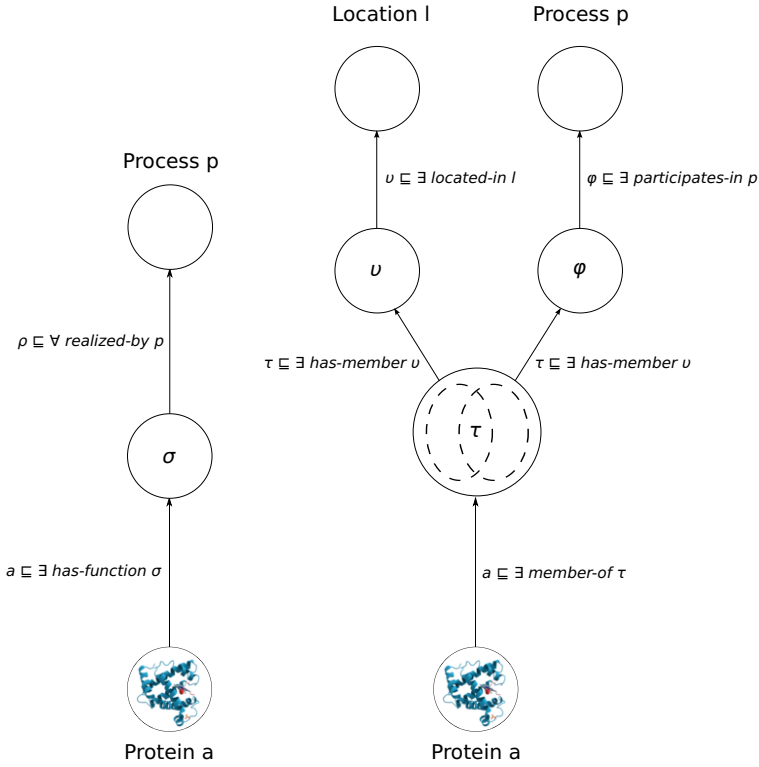

Figure 1. An overview over our approach. On the left side, the traditional representation pattern for representing possibilities in biomedical ontologies is shown, while on the right side of the figure, we illustrate our account. In our account, proteins are assumed to be a member of a timeless collection of proteins τ, all members of which are of the same type, and τ has some members which are actually participating in a process of type p, or alternative, actually being located at a location l. Our account falls in the OWL 2 EL profile while the traditional approach requires a more expressive logic.

Formally, given a class of proteins P, we generate the following protein-specific signature Σ_P:

- class P: a class whose instances are individual proteins of the type P;
- class P_{all}: a class whose instances are the set of all instances of P in the universe; P_{all} is intended to be a singleton class with exactly one instance;
- for each isoform P_i of P, a class P_i whose instances are individual proteins
- class $P_{generic}$: a class whose instances are proteins that have a shared evolutionary history, i.e., a group of *ortholog* proteins

Using this signature, we state the following axioms:

- P is a subclass of $P_{generic}$:
  - FOL: $\forall x(P(x) \rightarrow P_{generic}(x))$
  - OWL: $P \sqsubseteq P_{generic}$

  - both axioms are equivalent and fall in the OWL 2 EL profile

- protein isoforms are subclasses of the proteins of which they are an isoform:
  - $\forall x(P_i(x) \rightarrow P(x))$
  - $P_i \sqsubseteq P$
  - both axioms are equivalent and fall in the OWL 2 EL profile

- P_{all} is a singleton class:
  - FOL: $\exists x P_{all}(x) \wedge \forall x,y(P_{all}(x) \wedge P_{all}(y) \rightarrow x = y)$
  - OWL: $P_{all} \equiv \{p_{all}\}$, with p_{all} being a new constant symbol
  - the OWL representation falls in the OWL 2 EL profile and is satisfiable iff the FOL representation is satisfiable (proof sketch: the FOL representation is a direct consequence of the OWL equivalent class axiom; the $\rightarrow$ direction of the OWL axiom can be derived from the FOL axiom by Skolemization, the $\leftarrow$ direction then becomes a logical consequence).

- P_{all} is non-empty and has at least one member of type P:
  - FOL: $\forall x(P_{all}(x) \rightarrow \exists y(P(y) \wedge hasMember(x,y)))$
  - OWL: $P_{all} \sqsubseteq \exists hasMember.P$
  - both axioms are equivalent and fall in the OWL 2 EL profile

- P_{all} is homogenic, i.e., it has as members only proteins of type P:
  - FOL: $\forall x,y(P_{all}(x) \wedge hasMember(x,y) \rightarrow P(y))$
  - OWL: $P_{all} \sqsubseteq \forall hasMember.P$
  - both axioms are equivalent but do not fall in the OWL 2 EL profile

Furthermore, to assign a function realized by instances of F, and (possible) locations of type L, to a protein P, we use the following axioms:

- instances of P have the capability to perform processes F, restated as: some member of P_{all} (actually) participates in an instance of F:
  - FOL: $\forall x(P_{all}(x) \rightarrow \exists y,z(F(y) \wedge hasMember(x,z) \wedge participatesIn(z,y)))$
  - OWL: $P_{all} \sqsubseteq \exists hasMember.(\exists participatesIn.F)$
  - both axioms are equivalent and fall in the OWL 2 EL profile

- instances of P have the potential to be located at instances of L, restated as: some member of P_{all} is (actually) located at an instance of F:
  - FOL: $\forall x(P_{all}(x) \rightarrow \exists y,z(L(y) \wedge hasMember(x,z) \wedge locatedAt(z,y)))$
  - OWL: $P_{all} \sqsubseteq \exists hasMember.(\exists locatedAt.L)$
  - both axioms are equivalent and fall in the OWL 2 EL profile

To complete this account of proteins and their functions, we also have to provide a formal account of the relations we have introduced, i.e., of *participatesIn*, *locatedAt*, *hasMember* and its inverse *memberOf*. To maximize compatibility with many upper level ontologies and use cases, we only chose minimal axioms for each:

- *memberOf* is irreflexive and asymmetric
  - FOL: $\forall x(\neg memberOf(x,x)), \forall x,y(memberOf(x,y) \rightarrow \neg memberOf(y,x))$

  - both properties of *memberOf* can be stated equivalently in OWL 2 but neither property can be expressed within the OWL 2 EL profile. Additionally, the *memberOf* relation may further be assumed to satisfy the weak supplementation principle [37], an axiom that cannot be expressed in OWL.

- *participatesIn* is irreflexive and asymmetric
  - FOL: $\forall x(\neg participatesIn(x,x)), \forall x,y(participatesIn(x,y) \rightarrow \neg participatesIn(y,x))$
  - usually (except in 4D ontologies), the *participatesIn* relation is mainly restricted by the type of arguments it can take, the first being an endurant [38] (or continuant [39] or presential [40]), the second being a process [40]; irreflexivity and asymmetry then follow from the disjointness of these arguments.

Using these axioms, we can reformulate the notion that a protein of type P is capable of performing F (or having a function that is realized by processes of type F) using the following axiom:

$$P(x) \wedge \exists \zeta(hasFunction(x,\zeta) \wedge \forall f(realizes(f,\zeta) \rightarrow F(f))) \iff$$
$$P(x) \wedge \exists y(P_{all}(y) \wedge memberOf(x,y) \wedge \exists z(hasMember(y,z) \wedge$$
$$\exists f(participatesIn(z,f) \wedge F(f))))$$

There are several limitations of our representation, in particular with regard to their expressivity in OWL 2 EL. First, some axioms that relate proteins, and the collection of proteins to which they belong, are lost when we limit ourselves to the expressivity in OWL 2 EL. In particular, the axiom that asserts that the collection class P_{all} has only instances of P as member cannot be expressed in OWL 2 EL. Second, we do not use any temporal arguments in our relations to maintain compatibility with OWL in which relations can have at most two arguments. The collections containing all proteins of a particular type are "timeless" collections, i.e., they contain all instances of a type of protein across all times. If temporal arguments are required, for example when working within the framework of an upper level ontology in which relations contain temporal arguments, they can easily be added. In the simplest case, existential quantification over the temporal argument in the relations can be used to generate a representation equivalent to the one we use. Finally, we provide a (much) simplified account of *biological function*. To assign a biological function to all the instances of a type of protein, some instances of this type of protein must also be *observed*; additionally, not every type of participation in a process would be viewed as a function, but rather only particular types of participation [24,21,41]; these aspects are entirely ignored by our approach. In the future, this information may be added as a conservative extension of our theory.

6. Implementation: an Application to a Knowledge Base of Proteins

To evaluate whether our account of biological functions can be implemented on a large scale, we chose the UniProt database as our primary use case. UniProt [3] is one of the central knowledge bases used in modern biomedical science. It is a collaborative and integrated resource comprising a manually curated part of about half a million protein

sequences, known as SwissProt, and an automatically generated database, TrEMBL, with a rapidly growing 80 million sequences. Proteins in UniProt cover over 610,000 taxa [3].

UniProt has become one of the main authorities for the stable identification of proteins by providing UniProt accession numbers (which serve as unique identifiers for types of proteins). UniProt is a core service of the ELIXIR Infrastructure [42], other databases link UniProt records to diseases [43], phenotypes [44], sequence, tertiary structures, environments, etc., and UniProt itself contains links to more than 150 biological databases. UniProt, as a highly interlinked knowledge base, has been one of the first major biological databases to adopt the Linked Data [45] guidelines, provide RDF as a representation format as well as a public SPARQL endpoint to facilitate querying. Currently, as of 2016, UniProt comprises of almost 20 billion triples in RDF, and has frequently been used as benchmark for the performance of RDF stores and SPARQL queries [3].

The data in SwissProt, a part of UniProt, is manually curated from the literature and therefore presents a gold standard resource for sequence and function information on peptides and proteins. A key component of UniProt, and SwissProt, is the annotation of protein functions using the Gene Ontology (GO) [46].

The GO [8] is comprised of three ontologies (*molecular function*, *biological process* and *cellular component*) describing processes and cellular locations. The *cellular component* ontology is an ontology of cellular anatomy, while *molecular function* and *biological process* are ontologies of processes on different scales and levels of granularity. Despite its label, we assume that the classes contained in the *molecular function* ontology characterize processes, not *functions* or other types of entities that are fundamentally different from processes. While this may seem counterintuitive and different from some prior analyses of GO [47,48,49], recent versions of GO make it clear that the difference between *molecular function* and *biological process* is one of granularity, not between fundamentally different ontological categories. Consequently, both ontologies are no longer separated within GO but are integrated, with hundreds of mereological axioms (involving *part-of*, *has-part*) being asserted between classes from the *biological process* and *molecular function* ontologies [50]. The axioms related to these mereological relations within the GO are taken from the Basic Formal Ontology (BFO) [51] and the OBO Relationship Ontology [30], and based on the axioms provided within BFO (in particular the domain and range restrictions) it is clear that the *molecular function* ontology is an ontology of processes. For example, *catalytic activity* (`GO:0003824`), a class in the *molecular function* ontology, is used in an axiom (GO version 11 Feb 2016) that asserts it to be a subclass of $\exists partOf$.metabolic process, where *metabolic process* (`GO:0008152`) is a class in the *biological process* ontology. Similarly, another axiom involves *catalytic complex* (`GO:1902494`), from the *cellular component* ontology, and asserts *catalytic complex* as a subclass of $\exists hasFunction.(\forall realizedBy.$catalytic activity$)$ (using the *capable of* pattern); together with the range restrictions on *realizedBy* which forces its second argument to be an instance of *Process* (from BFO), this leads to the inference that *catalytic activity* is considered a process within GO.

We apply our method to the RDF representation of the SwissProt part of UniProt. In UniProt RDF, classes or types of proteins, such as *FOXP2*, are identified through an IRI, and the type of protein is associated with a class from the GO through a *classified-with* relation. For example, a triple

```
<O15409> up:classified-with go:GO_0048286
```

asserts that the protein with UniProt accession *O15409* (the human FOXP2 protein) is associated with *lung alveolus development* (GO:0048286). The same relation *classified-with* is used for all associations with the GO as well as for associations with keywords and some other databases.

We identify UniProt accessions with classes of proteins, and, according to our method for representing capabilities to perform in processes, automatically generate classes whose instance represents the collection of all proteins of that type, together with all the axioms in our theory that fall in the OWL 2 EL profile. The transformation is performed automatically for the full SwissProt. To determine whether a capability is the capability to perform a certain process or to be located at a particular location, we use the Elk reasoner [52] and query the GO for subclasses of *molecular function*, *biological process*, and *cellular component* respectively. If SwissProt associates a protein accession with a subclass of *cellular component*, we generate an axiom representing the capability of instances of the protein to be located at a location of the type specified by the GO class, otherwise we generate an axiom representing the capability of the protein to participate in a process of the type of the GO class.

Additionally, to explicitly identify the kind of species which can create a class of proteins, we use the information provided by UniProt about the taxon in which particular types of proteins are known to exist, and assert that each instance of a class of proteins was created in an instance of that taxon. For example, for the *FOXP2* protein in humans with UniProt accession O15409, we would add an axiom stating that each instance of *FOXP2* will have been created in an instance of *Human*, a class taken from the NCBI taxonomy [53] (it would be more precise to state that it has been created in something derived from an instance of *Human*, to also include proteins produced in human cell lines or transgenic organisms; however, we omit this aspect here and leave it as extension for future work).

We have implemented our conversion of UniProt to OWL in Jython, based on the Apache Jena API [54] for processing RDF files provided by UniProt and the OWL API [55] for generating the final ontology. The transformation is parallelized by processing each protein accession within SwissProt individually. The source code is freely available at https://github.com/bio-ontology-research-group/uniprot2owl.

The resulting ontology, automatically generated from SwissProt, is a complete representation of the proteins in SwissProt and their functions. The ontology is based on SwissProt, downloaded on 28 January 2016, and the GO and the NCBI Taxonomy downloaded on 15 February 2016. The resulting ontology contains 4,995,217 logical axioms, and 1,728,231 classes (excluding imported classes from the GO and NCBI Taxonomy), and can be downloaded from http://aber-owl.net/aber-owl/swissprot.owl.

7. Evaluation

We evaluate the resulting ontology both with respect to its capability for answering queries as well as with respect to the classification and query times. Specifically, our aim is to test whether classification and query time improves with our representation in comparison to the alternative approaches for representing possibilities that relies on the universal quantifier. For all tests, we use the Elk reasoner [56], an optimized reasoner

for the OWL 2 EL profile that supports parallel and incremental reasoning. We compare these results with the HermiT reasoner [57] which supports OWL 2 DL, and using a representation pattern for capabilities of proteins that relies on the universal quantifier. In particular, instead of our representation, we alternatively express the capability of a protein X to participate in P as $X \sqsubseteq \exists hasFunction.(\forall realizedBy.P)$ when using HermiT. The tests are performed using a workstation with an Intel Xeon E5-2680 with 128GB memory.

To classify the ontologies, we import the GO and the NCBI Taxonomy, and perform classification using the OWL API. Elk classifies the resulting ontology in 55 seconds, while HermiT fails to classify the ontology within one day.

We further verified whether we can answer certain queries against our knowledge base, and we test the performance of answering these queries. To answer queries, we use the Elk reasoner to identify all subclasses of an OWL class description. We use the class description $\exists memberof.(\exists hasMember.(\exists participatesIn.F))$ to query for capabilities to perform processes of type F, and the class description $\exists memberof.(\exists hasMember.(\exists locatedIn.L))$ to query for the capability to be located in L. To evaluate the query answering performance, we randomly select 1,000 classes from the GO and perform a query for each. The average answer time for these queries was 1.12 seconds. As HermiT could not classify the ontology when capabilities are alternatively represented using the universal quantifier, it could not answer any queries.

To further evaluate more complex queries, we also use another type of query in which we incorporate the taxon in which proteins have been created, and query for proteins with a particular capability *and* within a particular taxon, using the class description $\exists memberof.(\exists hasMember.(\exists participatesIn.F)) \sqcap \exists createdIn.T$ (and equivalently for locations) for 1,000 randomly selected pairs of classes from GO and the NCBI taxonomy. The average query time for one query was 1.15 seconds.

One of the main advantages of using OWL as a representation language for (parts of) UniProt is the increased capability to add axioms to the knowledge base in support of easier maintenance and quality control. For example, using OWL, it is easy to distinguish between cases where all proteins of some type have a function and where only some isoforms, or all evolutionarily related proteins, have this function; to state that every isoform of P has the capability to perform F, we would add the axiom $\exists isoformOf.P \sqsubseteq \exists memberOf.(\exists hasMember.(\exists participatesin.F))$.

Our ontology also provides a direct link between UniProt and manually created ontologies of proteins, in particular the Protein Ontology [58], which contain information about protein families, genes, sequences and modifications. Similarly to the Protein Ontology, our implementation leads to the possibility of using classes referring to proteins in complex class descriptions across biomedical ontologies, such as for diseases in which proteins are involved, or assays in which proteins are measured.

8. Discussion

Currently, the capability of a biological entity E to perform P is represented using the ontology design pattern $E \sqsubseteq \exists hasFunction.(\forall realizedBy.P)$, and basic reasoning tasks in knowledge representation languages supporting this pattern fall in complexity classes beyond polynomial time. Consequently, automated reasoning over functions in large

knowledge bases, such as those frequently found in biology and biomedicine, is challenging. Here, we have proposed a novel design pattern for expressing knowledge about biological functions, together with a biological and ontological justification, which can be expressed using the description logic EL++ [17] and implemented using the OWL 2 EL profile [16].

The resulting framework takes advantage of the use of OWL as a representation language rendering the increased ability to add axioms to the knowledge base in support of easier maintenance and quality control. It caters for the classification of proteins using the background knowledge of ontologies (for example the GO) and facilitates complex queries that combine information from the UniProt knowledge base with ontology knowledge. Moreover it leads to the possibility of using classes referring to proteins in complex class descriptions across biomedical ontologies, for example ontologies related to particular phenotypic manifestations, diseases etc. Ultimately, our framework is intended to sever as a model of how formal ontologies can be utilized for the improvement of biomedical knowledgebases as well as enhance our quality control abilities.

There are several limitations to our approach related primarily to the OWL 2 EL expressivity. For example, our approach does not include some axioms that relate proteins and the collection of proteins to which they belong, nor does it encompass temporal arguments. Crucially it provides a simplified account of *biological function*. As part of our future research, to aim to extend our theory to account for them.

9. Acknowledgements

The work was initiated at the 2015 DBCLS BioHackathon in Nagasaki with support from the Database Center for Life Science (DBCLS). This work was supported in parts by funding from the King Abdullah University of Science and Technology.

References

[1] H. Michalek. *A Formal Ontological Approach to Causality Embedded in the Top-Level Ontology of GFO*. PhD thesis, University of Leipzig, 2009. Forthcoming.

[2] L. Wright. Functions. *Philosophical Review*, 1973.

[3] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.

[4] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, 2014.

[5] G. Joshi-Tope et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(Database issue):D428–432, January 2005.

[6] A. Hatano, H. Chiba, H. A. Moesa, T. Taniguchi, S. Nagaie, K. Yamanegi, T. Takai-Igarashi, H. Tanaka, and W. Fujibuchi. Cellpedia: a repository for human cell information for cell studies and differentiation analyses. *Database*, 2011, 2011.

[7] N. Mitsuhashi, K. Fujieda, T. Tamura, S. Kawamoto, T. Takagi, and K. Okubo. BodyParts3D: 3D structure database for anatomical concepts. *Nucleic Acids Research*, 37(suppl 1):D782–D785, 2009.

[8] M. Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[9] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in Bioinformatics*, March 2015.

[10] S. Schulz, H. Stenzhorn, M. Boeker, and B. Smith. Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS – Electronic Journal in Communication, Information and Innovation in Health*, 3(1):31–45, 2009.

[11] S. Schulz, B. Suntisrivaraporn, F. Baader, and M. Boeker. SNOMED reaching its adolescence: Ontologists' and logicians' health check. *International Journal of Medical Informatics*, 78(Supplement 1):S86–S94, 2009.

[12] S. Schulz, K. Spackman, A. James, C. Cocos, and M. Boeker. Scalable representations of diseases in biomedical ontologies. *Journal of Biomedical Semantics*, 2(2):1–13, 2011.

[13] R. Hoehndorf, M. Dumontier, A. Oellrich, S. Wimalaratne, D. Rebholz-Schuhmann, P. Schofield, and G. V. Gkoutos. A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*, 27(7):1001–1008, April 2011.

[14] S. Jupp et al. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, 2014.

[15] B. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patelschneider, and U. Sattler. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309–322, November 2008.

[16] B. Motik, B. C. Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz. OWL 2 Web Ontology Language: Profiles. Recommendation, World Wide Web Consortium (W3C), 2009.

[17] F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathscr{EL}$ envelope. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

[18] P. Godfrey-Smith. Functions: Consensus without unity. *Pacific Philosophical Quarterly*, 74:196–208, 1993.

[19] R. G. Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press, 1988.

[20] J. R. Searle. *The Construction of Social Reality*. Free Press, January 1997.

[21] P. Diaz-Herrera. What is a biological function? In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2006)*, pages 128–140, 2006.

[22] M. Artiga. Re-organizing organizational accounts of function. *Applied Ontology*, 6(2):105–124, 2011.

[23] J. R. Searle. *The Construction of Social Reality*. Penguin Group, 1995.

[24] P. Burek. *Ontology of Functions*. PhD thesis, University of Leipzig, Institute of Informatics (IfI), 2006.

[25] D. P. Hill, B. Smith, M. S. McAndrews-Hill, and J. A. Blake. Gene ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9(5):1–9, 2008.

[26] de Angelis et al. Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. *Nat Genet*, 47(9):969–978, 2015.

[27] J. Röhl and L. Jansen. Why functions are not special dispositions: an improved classification of realizables for top-level ontologies. *Journal of Biomedical Semantics*, 5(1):1–16, 2014.

[28] R. Hoehndorf, A.-C. Ngonga Ngomo, and J. Kelso. Applying the functional abnormality ontology pattern to anatomical functions. *Journal of biomedical semantics*, 1(1):4+, March 2010.

[29] R. Arp and B. Smith. Function, role, and disposition in basic formal ontology. In *Proceedings of The 11th Annual Bio-Ontologies Meeting*, 2008.

[30] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biol*, 6(5):R46, 2005.

[31] R. Hoehndorf, A. Oellrich, M. Dumontier, J. Kelso, D. Rebholz-Schuhmann, and H. Herre. Relations as patterns: Bridging the gap between OBO and OWL. *BMC Bioinformatics*, 11(1):441+, 2010.

[32] F. Baader. *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press, January 2003.

[33] B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Modular reuse of ontologies: Theory and practice. *JAIR*, 31:273–318, 2008.

[34] C. D. Vescovo, P. Klinov, B. Parsia, U. Sattler, T. Schneider, and D. Tsarkov. Empirical study of logic-based modules: Cheap is cheerful. In *The 12th International Semantic Web Conference (ISWC2013)*, pages 81–96, 2013.

[35] F. Baader, C. Lutz, and B. Suntisrivaraporn. Efficient reasoning in $\mathscr{EL}^+$. In *Proceedings of the 2006 International Workshop on Description Logics (DL2006)*, CEUR-WS, 2006.

[36] C. Del Vescovo, D. D. Gessler, P. Klinov, B. Parsia, U. Sattler, T. Schneider, and A. Winget. Decomposition and modular structure of bioportal ontologies. In *The Semantic Web–ISWC 2011*, pages 130–145. Springer, 2011.

[37] G. Guizzardi. Representing collectives and their members in UML conceptual models: An ontological analysis. In J. Trujillo, G. Dobbie, H. Kangassalo, S. Hartmann, M. Kirchberg, M. Rossi, I. Reinhartz-

Berger, E. ZimÃąnyi, and F. Frasincar, editors, *ER Workshops*, volume 6413 of *Lecture Notes in Computer Science*, pages 265–274. Springer, 2010.

[38] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. WonderWeb Deliverable D18: Ontology library (final). Technical report, Laboratory for Applied Ontology – ISTC-CNR, Trento (Italy), 2003.

[39] P. Grenon. Spatio-temporality in basic formal ontology: Snap and span, upper-level ontology, and framework for formalization. IFOMIS Report 05/2003, University of Leipzig, Leipzig, 2003.

[40] H. Herre. General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In R. Poli, M. Healy, and A. Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, chapter 14, pages 297–345. Springer, Heidelberg, 2010.

[41] P. Burek, R. Hoehndorf, F. Loebe, J. Visagie, H. Herre, and J. Kelso. A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics*, 22(14):e66–73, July 2006.

[42] L. C. Crosswell and J. M. Thornton. ELIXIR: a distributed infrastructure for european biological data. *Trends in Biotechnology*, 30(5):241 – 242, 2012.

[43] J. Amberger, C. Bocchini, and A. Hamosh. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). *Hum Mutat*, 32:564–567, 2011.

[44] J. A. Blake, C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and G. Mouse Genome Database. The mouse genome database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res*, 42(Database issue):D810–7, 2014. DOI:10.1093/nar/gkt1225.

[45] C. Bizer, T. Heath, and T. Berners-Lee. Linked data—the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1âĂŞ22, 2009.

[46] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O'Donovan. The GOA database: Gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, 2015.

[47] B. Smith et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.

[48] B. Smith, J. Köhler, and A. Kumar. On the application of formal principles to life science data: A case study in the gene ontology. In *Proceedings of DILS 2004 (Data Integration in the Life Sciences)*, volume 2994 of *Lecture Notes in Bioinformatics*, pages 79–94, Berlin, 2004. Springer.

[49] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 609–613, 2003.

[50] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[51] R. Arp, B. Smith, and A. D. Spear. *Building Ontologies with Basic Formal Ontology*, volume 1. MIT Press, 2015.

[52] Y. Kazakov, M. Krötzsch, and F. Simančík. Unchain my $\mathscr{EL}$ reasoner. In *Proceedings of the 23rd International Workshop on Description Logics (DL'10)*, CEUR Workshop Proceedings. CEUR-WS.org, 2011.

[53] E. W. Sayers et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 37(suppl 1):D5–D15, 2009.

[54] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: Implementing the Semantic Web recommendations. Technical Report HPL-2003-146, Hewlett Packard, Bristol, UK, 2003.

[55] M. Horridge, S. Bechhofer, and O. Noppens. Igniting the OWL 1.1 touch paper: The OWL API. In *Proceedings of OWLED 2007: Third International Workshop on OWL Experiences and Directions*, 2007.

[56] Y. Kazakov, M. Krötzsch, and F. Simancik. The incredible ELK. *Journal of Automated Reasoning*, 53(1):1–61, 2014.

[57] B. Motik, R. Shearer, and I. Horrocks. Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.

[58] D. A. Natale et al. Protein ontology: a controlled structured network of protein entities. *Nucleic Acids Research*, 42(D1):D415–D421, 2014.