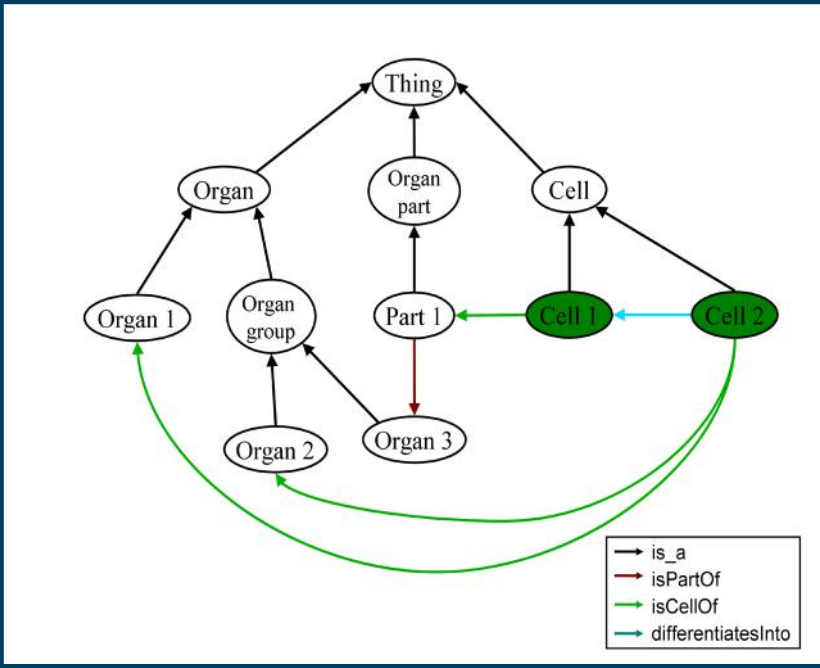
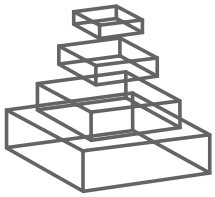
frontiers
RESEARCH TOPICS



BIOLOGICAL ONTOLOGIES AND SEMANTIC BIOLOGY

Topic Editor
John Hancock

frontiers in GENETICS



frontiers

ISSN 1664-8714
ISBN 978-2-88919-277-9
DOI 10.3389/978-2-88919-277-9

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

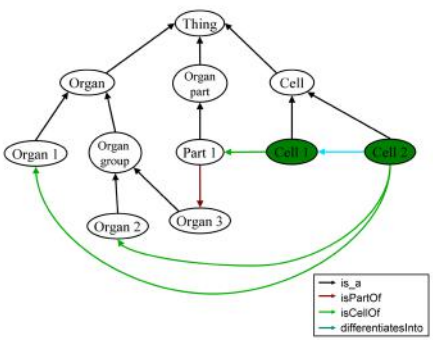
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

BIOLOGICAL ONTOLOGIES AND SEMANTIC BIOLOGY

Topic Editor:
John Hancock, University of Cambridge, United Kingdom



Abstracted and simplified view of the Cytomer ontology illustrating the handling of organs of an anatomical entity. The green nodes are the start nodes for the search function specified in the text. In this section, the entities are connected by four different relations given in the legend. Figure taken from: Dönitz J and Wingender E (2012) The ontology-based answers (OBA) service: a connector for embedded usage of ontologies in applications. *Front. Genet.* 3:197. doi: 10.3389/fgene.2012.00197

As the amount of biological information and its diversity accumulates massively there is a critical need to facilitate the integration of this data to allow new and unexpected conclusions to be drawn from it.

The Semantic Web is a new wave of web-based technologies that allows the linking of data between diverse data sets via standardised data formats ("big data"). Semantic Biology is the application of semantic web technology in the biological domain (including medical and health informatics). The Special Topic encompasses papers in this very broad area, including not only ontologies (development and applications), but also text mining, data integration and data analysis making use of the technologies of the Semantic Web.

Ontologies are a critical requirement for such integration as they allow conclusions drawn about biological experiments, or descriptions of biological entities, to be understandable and integratable despite being contained in different databases and analysed by different software systems. Ontologies are the standard structures used in biology, and more broadly in computer science, to hold standardized terminologies for particular domains of knowledge. Ontologies consist of sets of standard terms, which are defined and may have synonyms for ease of searching and to accommodate different usages by different communities. These terms are linked by standard relationships, such as "is_a" (an eye "is_a" sense organ) or "part_of" (an eye is "part_of" a head). By linking terms in this way, more detailed, or granular, terms can be linked to broader terms, allowing computation to be carried out that takes these relationships into account.

Table of Contents

04 ***Editorial: Biological Ontologies and Semantic Biology***
John M. Hancock

06 ***The AEO, an Ontology of Anatomical Entities for Classifying Animal Tissues and Organs***
Jonathan B. L. Bard

13 ***IMGT-Ontology 2012***
Véronique Giudicelli and Marie-Paule Lefranc

29 ***Three Ontologies to Define Phenotype Measurement Data***
Mary Shimoyama, Rajni Nigam, Leslie Sanders Mcintosh, Rakesh Nagarajan, Treva Rice, D. C. Rao and Melinda R. Dwinell

39 ***Development and Use of Ontologies Inside the Neuroscience Information Framework: A Practical Approach***
Fahim T. Imam, Stephen D. Larson, Anita Bandrowski, Jeffery S. Grethe, Amarnath Gupta and Maryann E. Martone

51 ***An Ontological Analysis of Some Biological Ontologies***
Briti Deb

54 ***The Choice Between Mapman and Gene Ontology for Automated Gene Function Prediction in Plant Science***
Sebastian Klie and Zoran Nikoloski

68 ***Use of the Protein Ontology for Multi-Faceted Analysis of Biological Processes: A Case Study of the Spindle Checkpoint***
Karen E. Ross, Cecilia N. Arighi, Jia Ren, Darren A. Natale, Hongzhan Huang and Cathy H. Wu

83 ***Annotation Extension Through Protein Family Annotation Coherence Metrics***
Hugo P. Bastos, Luka A. Clarke and Francisco M. Couto

93 ***The Ontology-Based Answers (OBA) Service: A Connector for Embedded Usage of Ontologies in Applications***
Jürgen Dönitz and Edgar Wingender

104 ***Social Networks for Ehealth Solutions on Cloud***
Briti Deb and Satish N. Srirama

frontiers in
GENETICS

EDITORIAL
published: 04 February 2014
doi: 10.3389/fgene.2014.00018

Editorial: biological ontologies and semantic biology

John M. Hancock*

Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK
**Correspondence: jmhancock@gmail.com*

Edited and Reviewed by:
Richard D. Emes, University of Nottingham, UK

Keywords: semantic biology, biological ontologies, semantic web, data representation, data analysis

As the amount of biological data and its diversity accumulates massively there is a critical need to facilitate the integration of this data to allow new and unexpected conclusions to be drawn from it.

The Semantic Web comprises web-based technologies that allow linking of data between diverse data sets. Semantic Biology is the application of semantic web technology in the biological domain (including medical and health informatics). The Special Topic in Biological Ontologies and Semantic Biology brings together papers in this broad area—which spans computer science, computational biology and bioinformatics—providing a platform for strengthening what is still a new and underappreciated area of research.

A key aspect of semantic biology is the description of biological, and biology-related, entities using ontologies. Ontologies are a critical requirement for such integration as they allow conclusions drawn about biological experiments, or descriptions of biological entities, to be understandable and integratable despite being contained in different databases and analyzed by different software systems. Ontologies are the standard structures used in biology, and more broadly in computer science, to hold standard terminologies for particular domains of knowledge. They consist of sets of standard terms, which are defined and may have synonyms for ease of searching and to accommodate different usages by different communities. These terms are linked by standard relationships, such as "is_a" (an eye "is_a" sense organ) or "part_of" (an eye is "part_of" a head). In this way more detailed (granular) terms can be linked to broader terms, allowing computation to be carried out that takes these relationships into account.

The classical biological ontology is the Gene Ontology (GO) (Ashburner et al., 2000) which addresses aspects of gene function, the processes in which they participate and the localization of gene products. Increasingly, semantic biology requires the linkage of these concepts to other biological features. Three such biological entities are included in the Special Topic. The Anatomical Entity Ontology (AEO) (Bard, 2012) provides a typology of anatomical entities across species that is linked to cell types (via links to the cell ontology). Amongst others things, this allows linkage of anatomical structures across species, allowing inferences of homology and comparison of features such as gene and protein expression across species.

Another cross-species ontology, and one that complements work on anatomy, is described by Giudicelli and Lefranc (2012). They provide an update on the IMGT-Ontology which is an ontology of immunogenetics and immunoinformatics used in the international ImMunoGeneTics information system® (http://www.imgt.org). The IMGT-Ontology describes a range of immunogenetics concepts (immunoglobulins or antibodies, T cell receptors, major histocompatibility (MH) proteins of humans and other vertebrates, proteins of the immunoglobulin superfamily and MH superfamily, related proteins of the immune system of vertebrates and invertebrates, therapeutic monoclonal antibodies, fusion proteins for immune applications, and composite proteins for clinical applications).

A key problem for semantic biology is linking data on phenotypic measurements between model organisms, used to understand human disease, and clinical observations made in humans. This has been an active area of research in recent years (Hancock et al., 2009; Schofield et al., 2010). Shimoyama et al. (2012) make an important contribution to this area by describing a set of ontologies used to describe clinical measurements, measurement methods and experimental conditions for traits common to rat and man (and, by extension, in other mammalian model systems such as mouse and, potentially, more distantly related species). These measurements are similar to those used in large-scale phenotyping experiments (Hancock and Gates, 2011) so that this ontology system provides a potentially valuable mechanism for the study of genotype-phenotype relations in mammals.

Going beyond the underlying ontological structures used to describe biological data Imam et al. (2012) describe an integrated set of ontologies used within the Neuroscience Information Framework (www.neuinfo.org/), which describe major domains in neuroscience, including diseases, brain anatomy, cell types, sub-cellular anatomy, small molecules, techniques, and resource descriptors. This application provides a valuable insight into how sets of existing ontologies can be integrated with novel, more application-specific ontologies and structures to underpin a semantic-based knowledge system. NIF links logically consistent sets of terms into single structures but forms links between these logically consistent sets using bridging modules. Deb (2012) argues for an alternative approach using a single upper level (foundational) ontology to link specific biological domain ontologies.

A key issue that any such framework raises is how to compare and choose appropriate ontologies for any given system. A typical default position in biological applications is to accept the ontologies held in the open biological ontologies set (Smith et al., 2007). Here Klie and Nikoloski (2012) argue that ontology choice is to a degree application-specific and that domain-specific ontologies may in some cases be more useful than general ontologies such as the GO.

The major purpose of developing biological ontologies (rather than simpler controlled vocabularies) is to make use of the relations implicit in ontologies to facilitate analysis and annotation. These topics are addressed by two papers in this series. Ross et al. (2013) describe the use of the PRotein Ontology to carry out cross-species comparisons of function in the spindle checkpoint pathway. Bastos et al. (2013) consider the use of subsets of functionally coherent proteins to improve functional annotation in a protein family.

Finally, advances in technology provide new opportunities for the use of semantically-enriched data in applications that are only minimally ontology-aware. Dönitz and Wingender (2012) describe a web-based service that can be accessed from any application to make use of standard ontologies, removing a significant burden to application development. At a higher level, Deb and Srirama (2013) provide us with a view of how the data and ontologies currently being produced might be linked and accessed via cloud infrastructures and describe some of the problems this raises in the domain of human eHealth.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bard, J. B. L. (2012). The AEO, an ontology of anatomical entities for classifying animal tissues and organs. *Front. Genet.* 3:18. doi: 10.3389/fgene.2012.00018

Bastos, H. P., Clarke, L. A., and Couto, F. M. (2013). Annotation extension through protein family annotation coherence metrics. *Front. Genet.* 4:201. doi: 10.3389/fgene.2013.00201

Deb, B. (2012). An ontological analysis of some biological ontologies. *Front. Genet.* 3:269. doi: 10.3389/fgene.2012.00269

Deb, B., and Srirama, S. N. (2013). Social networks for eHealth solutions on cloud. *Front. Genet.* 4:171. doi: 10.3389/fgene.2013.00171

Dönitz, J., and Wingender, E. (2012). The ontology-based answers (OBA) service: a connector for embedded usage of ontologies in applications. *Front. Genet.* 3:197. doi: 10.3389/fgene.2012.00197

Giudicelli, V., and Lefranc, M. P. (2012). Imgt-ontology 2012. *Front. Genet.* 3:79. doi: 10.3389/fgene.2012.00079

Hancock, J. M., and Gates, H. (2011). "The informatics of high-throughput mouse phenotyping: EUMODIC and beyond," in *Mouse as a Model Organism–From Animals to Cells*, eds *C. Brakebusch and T. Pihlajaniemi (Berlin: Springer),* 77–88.

Hancock, J. M., Mallon, A. M., Beck, T., Gkoutos, G. V., Mungall, C., and Schofield, P. N. (2009). Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm. Genome* 20, 457–461. doi: 10.1007/s00335-009-9208-3

Imam, F. T., Larson, S. D., Bandrowski, A., Grethe, J. S., Gupta, A., and Martone, M. E. (2012). Development and use of ontologies inside the neuroscience information framework: a practical approach. *Front. Genet.* 3:111. doi: 10.3389/fgene.2012.00111

Klie, S., and Nikoloski, Z. (2012). The choice between mapman and gene ontology for automated gene function prediction in plant science. *Front. Genet.* 3:115. doi: 10.3389/fgene.2012.00115

Ross, K. E., Arighi, C. N., Ren, J., Natale, D. A., Huang, H., and Wu, C. H. (2013). Use of the protein ontology for multi-faceted analysis of biological processes: a case study of the spindle checkpoint. *Front. Genet.* 4:62. doi: 10.3389/fgene.2013.00062

Schofield, P. N., Gkoutos, G. V., Gruenberger, M., Sundberg, J. P., and Hancock, J. M. (2010). Phenotype ontologies for mouse and man; bridging the semantic gap. *Dis. Model. Mech.* 3, 281–289. doi: 10.1242/dmm.002790

Shimoyama, M., Nigam, R., McIntosh, L. S., Nagarajan, R., Rice, T., Rao, D. C., et al. (2012). Three ontologies to define phenotype measurement data. *Front. Genet.* 3:87. doi: 10.3389/fgene.2012.00087

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255. doi: 10.1038/nbt1346

Received: 09 January 2014; accepted: 21 January 2014; published online: 04 February 2014.

Citation: Hancock JM (2014) Editorial: biological ontologies and semantic biology. Front. Genet. ***5****:18. doi: 10.3389/fgene.2014.00018*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 14 February 2012
doi: 10.3389/fgene.2012.00018

The AEO, an ontology of anatomical entities for classifying animal tissues and organs

Jonathan B. L. Bard*

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK

Edited by:
John Hancock, Medical Research Council, UK

Reviewed by:
Gaurav Sablok, Huazhong Agricultural University, China
Qiangfeng Cliff Zhang, Columbia University, USA
David Osumi-Sutherland, Information Technology and Services, UK
Paula Mabee, University of South Dakota, USA

***Correspondence:**
Jonathan B. L. Bard, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK.
e-mail: j.bard@ed.ac.uk

This paper describes the AEO, an ontology of anatomical entities that expands the common anatomy reference ontology (CARO) and whose major novel feature is a type hierarchy of ∼160 anatomical terms. The breadth of the AEO is wider than CARO as it includes both developmental and gender-specific classes, while the granularity of the AEO terms is at a level adequate to classify simple-tissues (∼70 classes) characterized by their containing a predominantly single cell-type. For convenience and to facilitate interoperability, the AEO contains an abbreviated version of the ontology of cell-types (∼100 classes) that is linked to these simple-tissue types. The AEO was initially based on an analysis of a broad range of animal anatomy ontologies and then upgraded as it was used to classify the ∼2500 concepts in a new version of the ontology of human developmental anatomy (www.obofoundry.org/), a process that led to significant improvements in its structure and content, albeit with a possible focus on mammalian embryos. The AEO is intended to provide the formal classification expected in contemporary ontologies as well as capturing knowledge about anatomical structures not currently included in anatomical ontologies. The AEO may thus be useful in increasing the amount of tissue and cell-type knowledge in other anatomy ontologies, facilitating annotation of tissues that share common features, and enabling interoperability across anatomy ontologies. The AEO can be downloaded from http://www.obofoundry.org/.

Keywords: anatomical hierarchy, cell-type assignations, ontology, tissue classification

INTRODUCTION

Formal anatomical ontologies are now an important component of the informatics infrastructure of model organism and other databases (Bard, 2008; for a review of anatomy ontologies, see the papers in Burger et al., 2008; for examples, see[1]) and are also a key part of the informatics tools intended to explore biomedical databases. These ontologies primarily use *part_of* as their main structural relationship (e.g., every heart is *part_of* a cardiovascular system) because the smaller anatomical entities (usually referred to as tissues) are naturally seen as the constituent parts of larger ones, albeit that one tissue may be a part of more than one anatomical system (e.g., the *femur* is *part_of* the *lower limb* and the *skeletal system*). In addition, this relation is particularly important within database schemas for querying such tissue-associated knowledge as gene-expression data (e.g., the totality of the genes expressed in the heart at some developmental stage is the sum of the genes expressed in its parts).

In addition to *part_of* relationships, anatomical ontologies also need a classification or type hierarchy in which every term is related by an *is_a* or type relationship to a higher class term (e.g., the *femur is_a bone*, the *deltoid is_a muscle*). This relationship is required for three reasons: first, to ground the ontology within a standard formal structure (ontologies are based on classes within superclasses); second, many ontology visualization tools require this relationship; and third, this classification assigns to a term anatomical knowledge that would otherwise be missing.

An informal way of handling this issue is to indicate tissue type within an anatomy ontology through the use of high-level terms (e.g., leg skeleton, limb muscle system, cranial ganglia) but, while this is sometimes adequate for navigation around the ontology, it cannot be viewed as satisfactory or rigorous because it is based on a part_of rather than an is_a or type relationship. A better approach has been to use the common anatomy reference ontology (CARO) to classify anatomical structures (Haendel et al., 2008). This very high-level ontology of anatomical types is intended to provide a coarse framework of low-granularity for referencing the tissues of adult organism on the basis of anatomical level. Its 80 or so terms cover all anatomical classes from a hermaphroditic organism to an epithelium's basal lamina; only about 16 of them, however, can be used to classify tissues and cell-types (e.g., *organ system, compound organ, multi-tissue structure*). The only histological classification in the CARO covers the different types of epithelia; no other tissues (e.g., neuronal, muscular, and mesenchymal) merit a mention.

While the CARO provides a high-level class for a structure of any scale and so can be used to satisfy the requirement that every class have a superclass, its very low-granularity means that it can only annotate the thousands of tissue types that are known with very limited knowledge about anatomical structure. The restrictions of the CARO have been informally discussed within the field

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

for some time and additions are beginning to be made. Thus the curator of the *Drosophila* anatomy ontology needed to add a few new type terms (e.g., *row*) for classifying adult fly tissues. More recently, the vertebrate musculoskeletal anatomy ontology (namespace: VAO) has been produced (see text footnote 1), also using the CARO for its high-level terms, and this ontology meets the need for a new and much richer set of classes for this subset of anatomy. A more serious omission in CARO is that, because it was designed for adult anatomies, it lacks terms for developing tissues, a major focus of many anatomical ontologies. These and other class terms have been included in Uberon (Washington et al., 2009), an integrated cross-species ontology with high-level CARO terms and classified by structure, function, and developmental lineage, but not in any detail by tissue type. It is thus clear that an ontology for anatomy tissues that is both richer and finer-grained than CARO is required if one wishes to include structural knowledge about tissues in anatomy ontologies.

This paper describes the ontology of anatomical entities (AEO), an expansion of the CARO. The AEO is intended to capture and classify knowledge about anatomical structures not currently included in anatomical ontologies and includes ∼100 new classes structured using the *is_a* relationship. The AEO terms were selected partly through analysis of histology and anatomy books, partly through logical analysis, partly for their use in classifying the new ontology of human developmental anatomy (∼2500 terms) and partly through examination of a range of animal ontologies (whose ids are included where appropriate). The granularity of the AEO terms is at a level adequate for tissues of a predominantly single cell-type and, and these are given through *has_part* relationships to an abbreviated version of the ontology of cell-types (∼90 classes) included in the AEO. The AEO may be useful in increasing the amount of tissue and cell-type knowledge in other anatomy ontologies, facilitating annotation of tissues that share common features, and enabling interoperability across anatomy ontologies.

MATERIALS AND METHODS

The AEO uses the CARO as its basis for high-level classes. Terms for the histological information used to link cell-types to tissues came from standard textbooks (e.g., Ross et al., 1995; Standring, 2008; human anatomy is, for obvious reasons, analyzed in far greater depth than that of other organisms). Additional terms came from an analysis of other adult and anatomical ontologies from the biomedical ontologies site, particularly the VAO and, in these cases, the original ids are stored as dbxrefs. All ontologies mentioned in the paper are available from the OBO foundry (see text footnote 1). In this context, it might have been appropriate to incorporate within the AEO the terms and the structure of the VAO. The major skeletal terms from the VAO have been included (with definitions and ids), but the structure of the VAO was not used, mainly because it is much larger, more complex, and more fine-grained than is appropriate for the AEO and partly because some of the finer details of classification is at odds with expectation.

The process of constructing the AEO is described below. In brief, a first draft was made on the basis of inspection of a wide range of anatomical ontologies combined with general reading. This was used to classify the ontology of human developmental anatomy which has ∼2500 concepts. This process exposed weaknesses and omissions that were successively corrected.

Because the granularity of the AEO is designed to include anatomical entities of a single cell-type (*simple-tissue* or its synonym *portion of tissue*), it seemed sensible to include these cell-types within the ontology. While this could have been done using dbxrefs to the cell-type ontology, it seemed more appropriate to include the cell-type terms within the ontology so that a partonomy relationship could be assigned. A subset of the cell-type ontology was therefore included within the AEO and its terms linked to appropriate simple-tissue via the *has_part* relationship which carries the meaning that tissue A includes within it at least some of cell-type B.

The AEO terms not originally present in the CARO carry AEO ids whose numbers do not overlap with CARO ids (see Discussion) and is authored in the obo format[2] using the OBO-Edit[3] (Day-Richter et al., 2007.) and CoBrA[4] (Aitken et al., 2005) browsers (the former for complex ontologies, the latter for simple ones). Terms also carry appropriate dbxrefs from the *Drosophila*, VAO, zebrafish, Uberon, and human developmental anatomy ontologies. Obo-Edit includes the ability to make *disjoint_from* links that facilitate inconsistency checking (Rector, 2003) and such links have been made for *male* and *female anatomical structures*, and for *material* and *immaterial anatomical structures*.

The obo ontology is available from the OBO foundry (see text footnote 1). For Protégé users, the OWL version is generated automatically by the OBO Foundry pipeline, and is available from the same URL.

RESULTS

DESIGN FEATURES

The key aim of the AEO was to provide at least one unambiguous type term for every tissue in the anatomical ontology for an animal, whether adult, or developing. This turned out to be a more complicated process than originally expected, and what is described below is the final result of a series of iterations as drafts of the AEO were used for annotation (see below). The initial stage in making the AEO involved making a series of choices.

The first decision resulted from considering whether further high-level terms were needed in the CARO, and two omissions were noted: the exclusion of *gender-specific* and *embryonic anatomical entities*. The former was straightforward to add, but the latter, important for anatomy ontologies that cover developing organisms, was more difficult. The problem in choosing subterms here lies in the fact that all tissues in an embryo are developing tissues (even if they are fully functional and just growing, e.g., the late metanephros) and there is little point in annotating every term in an ontology with *is_a developing tissue*. As a result, a minimalist view was taken here and the terms in the *developing tissue* branch of the ontology were limited to those that were likely to be populated, were not present in an adult organism and had a useful developmental implication (**Figures 1** and **2**). Excluded from the list are any terms that imply lineage (such as may be found in

[2]http://purl.obolibrary.org/obo/oboformat/spec.html
[3]http://oboedit.org
[4]http://www.xspan.org/cobra/index.html

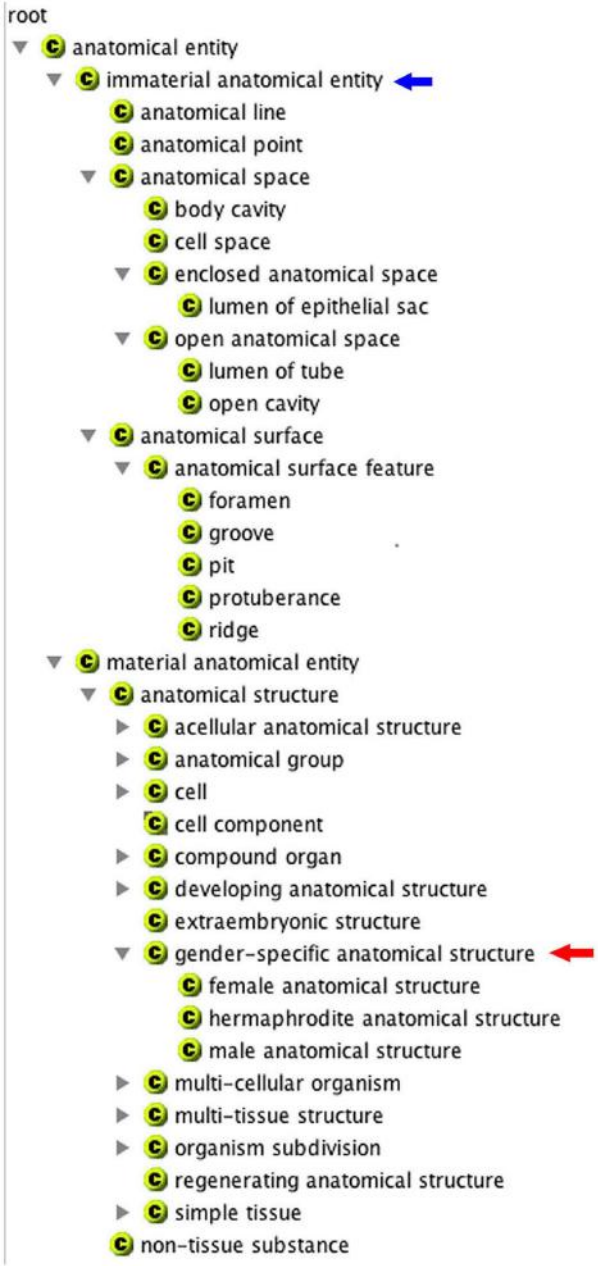


FIGURE 1 | The AEO shown in the COBrA browser. Here, the hierarchies for *immaterial anatomical entities* (blue arrow) and *gender-specific anatomical entities* (red arrow) are expanded.

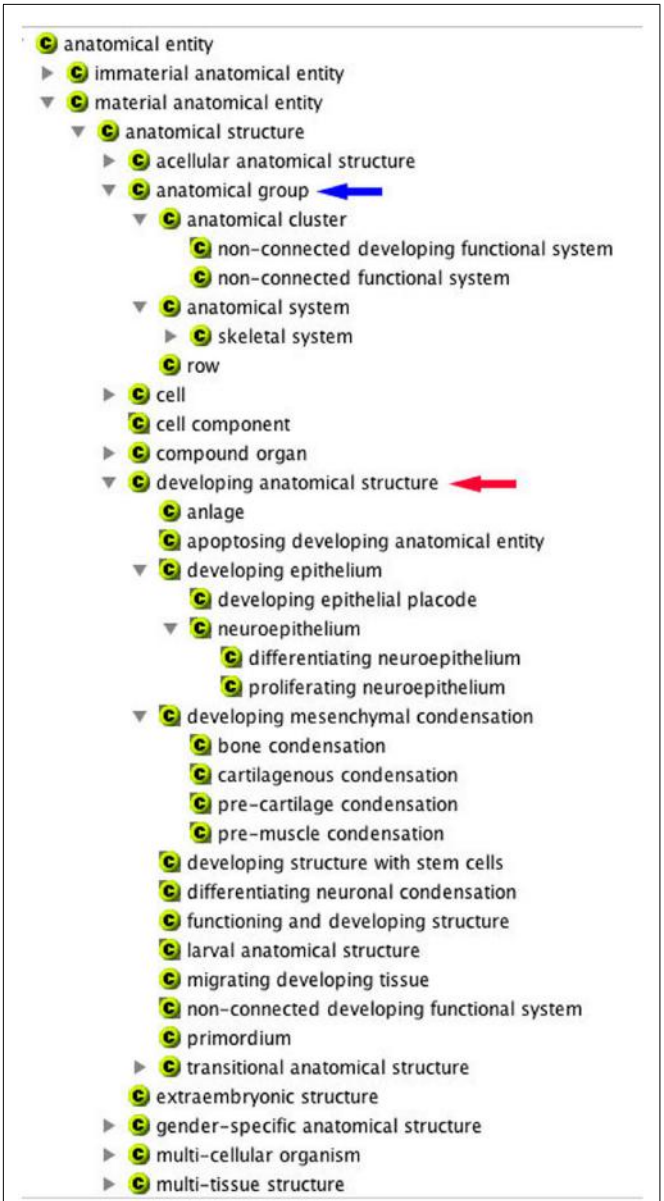


FIGURE 2 | The AEO shown in the in the COBrA browser. Here, the hierarchies for *anatomical group* (blue arrow) and *developing anatomical structure* (red arrow) are expanded.

Uberon); this is mainly because there are few if any tight lineage restrictions on tissue morphology. The list is of developmental classes is thus short and may need to be extended.

The second decision focused on the depth of the ontology, and here a CARO definition proved key: the CARO defines a *portion of tissue* as "anatomical structure that consists of similar cells and intercellular matrix, aggregated according to genetically determined spatial relationships." This definition fits comfortably with an anatomist's view of the simplest tissue by implying that it has a defined boundary and has cells predominantly of a single class (although this definition does raise the occasional problem – see below). One advantage of going down to this simple level of structure was that it enabled each leaf term to be annotated with its cell-types (as detailed in the cell-type ontology).

The third decision in making the AEO lay in choosing the breadth of the hierarchy. The coverage should be good enough to be useful without being overwhelmingly detailed, and anatomists have produced very detailed catalogs of tissue classes: Gray's anatomy (Standring, 2008), for example, lists >8 types of joint, most of which are rare. In making the AEO, all the major animal ontologies (i.e., plant and fungal ontologies are excluded) available

at the OBO library (see text footnote 1) were examined, and terms were chosen on the basis that they were likely to be useful (i.e., populated) and clear in meaning to anatomists. Thus, only the two most common classes of joint (*synovial* and *fibrous joints*) are included in the AEO as specific subclasses of joint; the former *is_a multi-tissue structure* and the latter *is_a* a *simple-tissue*, while cartilage is not subdivided. Also excluded are accessory bones (a subclass of sesamoid bones), bursas (a subclass of epithelial sac), and venules and arterioles (because they are all both unnamed and dispersed). Because skeletal terms are so common and useful to anatomists and to evolutionary biologists, it seemed sensible to group them all as parts under a new term *skeletal system*, a subclass of *anatomical system*.

There is a further small point: the terms of the AEO are intended to be clear in meaning to any biologist: as the ontology is intended for experimentalists who wish to annotate terms and access data, it is therefore important that the terms be those in common use. In this context, no anatomist has an intuitive sense of what the CARO term *portion of tissue* means, so the AEO uses that tem as a synonym for its replacement *simple-tissue* (similarly, the term *portion of organism substance* has been made a synonym of the more intuitively obvious term *non-tissue substance*).

MAKING THE ONTOLOGY

The major structural additions to the CARO that seemed necessary beyond adding *developmental* and *gender-specific tissues* were the expansion of some top-level class terms such as *immaterial anatomical entity*, *anatomical groups,* and *organism subdivision*, and here it seemed sensible to include the obvious major categories (*head*, *body*, etc., see below). Similarly, the class *multi-tissue structure* was felt to be too broad and in need of subterms, perhaps the most important of which is *tissues with stem cells*.

The class of *immaterial anatomical entities* (i.e., terms that refer to features rather than tissues) is treated lightly in the CARO: its few terms merely specify dimension (*anatomical line*, *point*, *surface,* and *space*). This terseness does less than justice to the richness of surfaces and volumes in organisms so the AEO includes several more terms (**Figure 1**) that can be used to group immaterial entities with common topological features (e.g., *open* and *enclosed cavities*, **Figure 1**). One interesting question here concerned how to class *surface pits* and *grooves* (e.g., the *otic pit*): should they be viewed as anatomical spaces (3D) or as surface features? Perhaps the most logical way to handle this would be to view the cells bounding the feature as a simple-tissue and the enclosed space (with a virtual enclosing surface) as an immaterial anatomic space. This would mean distinguishing between, say, the otic pit space and the otic pit epithelium, but standard anatomical usage implies that the *otic pit* is actually a surface feature within the surface epithelium. After some thought, the latter option was chosen with the user having the further option of annotating the term with a tissue type, so allowing both the cell-type and the geometric feature to be captured. Should a user specifically wish to refer to the space within the pit, the volume can be classified as a *lumen of an epithelial sac*. There is, it should be said, some vagueness in saying that an entity can be both a material and an immaterial entity; the values in doing this are terseness and the ability to captures some sense of tissue geometry, the price is the risk, albeit small, of ambiguity.

The other key task was the choice of simple-tissue leaf terms and this was mainly done on the basis of analyzing anatomy ontologies and histology texts. The net result was a major expansion in the CARO class *simple-tissue* (*portion of tissue*) which now has eight subclasses rather than one, with these subclasses opening up to two further levels which cover a further 60 or so classes (**Figure 3**). One anomalous term that has been included under *neuronal tissue* is *nerve fiber tract*: even though such tracts are composed of axons rather than of complete cells and so are not a tissue in the normal meaning of the word, this term was included because *nerve fiber tracts* are both named and important. As neither the CARO nor the cell-type ontology has a natural class that includes anatomical entities composed of cell parts, the GO definition for *neuron projection bundle* (and GO id dbxref) has been used here (and the synonym included). In a sense, all neuronal tissues are anomalous because the cell bodies and axons are not found within the same structure and it would have seemed odd to have included *nerve fiber tract* under any heading other than *simple-tissue*.

As a result of this, a draft extension to the CARO was constructed with ∼70 new terms.

IMPROVING DRAFT VERSIONS OF THE AEO

The AEO is intended to provide an *is_a* link for any anatomical concept. As the initial draft was based on inspection of a range of anatomical ontologies for animals, it met this criterion for most animal tissues. A harsher and finer granularity test was its ability to provide type terms for all the concepts in a detailed anatomical ontology. For this, drafts of the AEO were used to provide an obvious type term for the ∼2500 tissues in the new and integrated ontology of human developmental anatomy (namespace: EHDAA2; current draft available from http://www.obofoundry.org/) which is currently being constructed by the author from one made a decade ago (Hunter et al., 2003) that included a separate ontology for each Carnegie stage (1–20). The process of annotating a very wide range of anatomical classes from major organ systems down to simple-tissues in EHDAA2 identified inadequacies in draft AEO ontologies and required many changes to both the terms and the structure of the AEO. The introduction of *developing anatomical structure* and *gender-specific embryological structure* has already been mentioned (**Figures 1** and **2**). Another example was the amplification of *organism subdivision*. This last category proved useful, for example, in grouping the many and disparate entities within the *head* using *part_of* relationships (**Figure 5**). As things currently stand, there is at least one easily assignable class term for all anatomical terms so far examined, be it a leaf node (e.g., *metanephric mesenchyme is_a developing mesenchymal condensation*) or a higher level concept (*somite group is_a row*).

During this exercise, another ∼30 terms were added. In the current version, 13 of the new terms are classes of *immaterial anatomical entity*, ∼20 are *developing anatomical entities*, ∼40 are new types of simple-tissue, 15 are *multi-tissue structures*, seven are *anatomical groups,* and a few others are distributed under various headings. Further terms can easily be added if users feel that they would be needed. The current version of the AEO thus includes ∼160 anatomical classes and ∼100 cell-types.

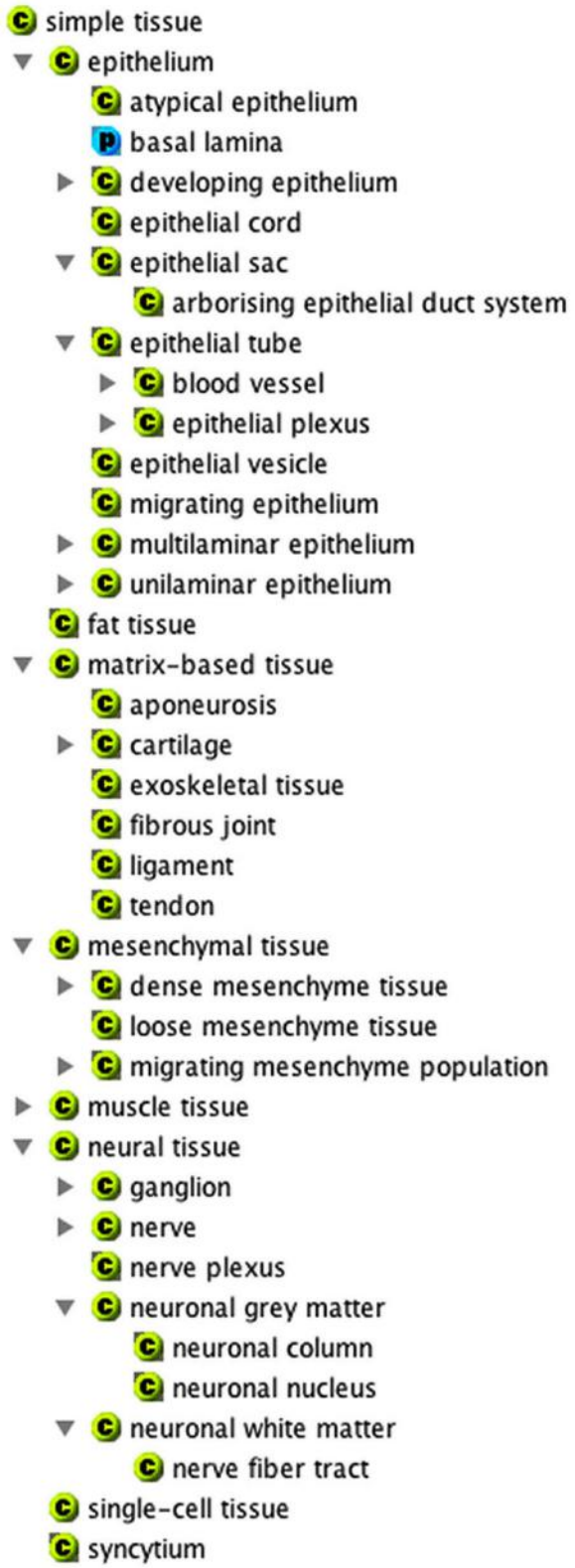


FIGURE 3 | The *simple-tissue hierarchy* of the AEO shown in the COBrA browser. All top and secondary levels terms are shown together with some tertiary level ones.

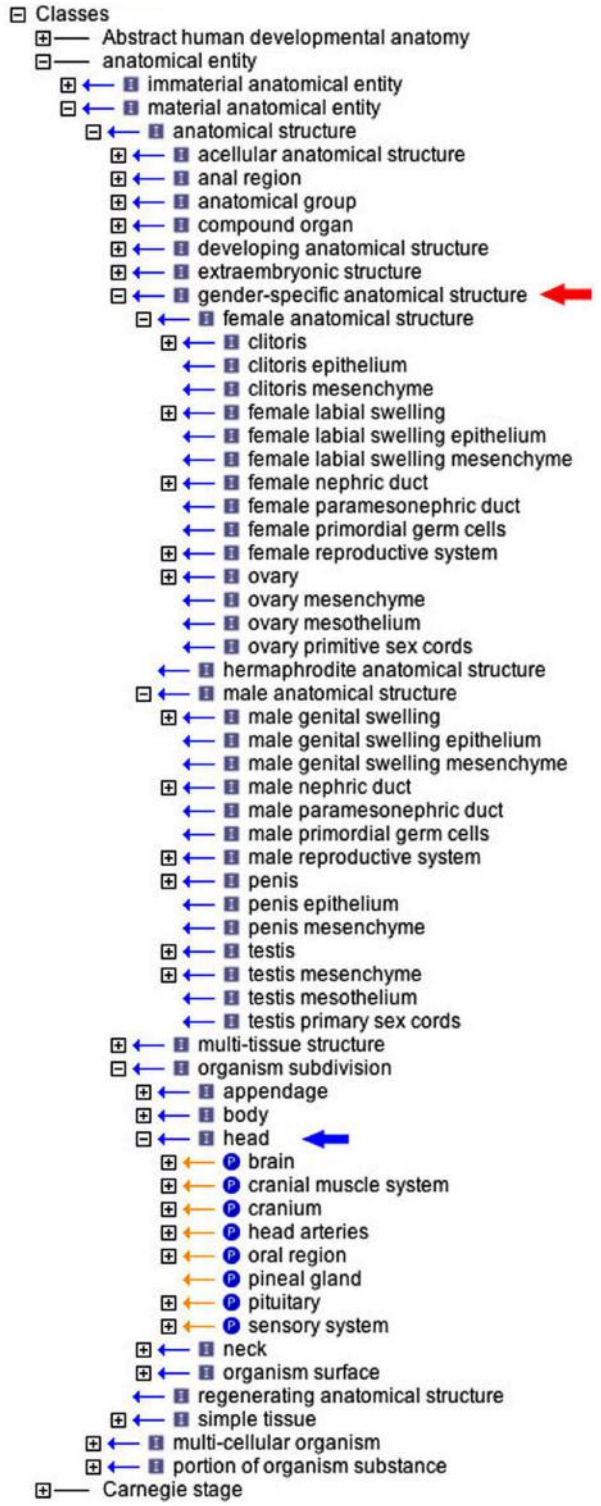


FIGURE 4 | The use of the AEO in classifying the ontology of human developmental anatomy (EHDAA2). The ontology is opened in the OBO-Edit browser to demonstrate (i) those tissues classed (*is_a* relationship) within the *gender-specific anatomical structure* hierarchy (red arrow), and (ii) the *head* category of *organism subdivision* with its constituent organ groups (*part_of* relationship).

Table 1 | The AEO obo file entry for *autonomic ganglion*.

(Term)
ID: AEO:0001001
Name: autonomic ganglion
Namespace: anatomical_entity_ontology
Def: "a ganglion that is part of the autonomic nervous system." (JB:AEO)
Is_a: AEO:0000135! ganglion
Relationship: has_part CL:0000107! autonomic neuron
Relationship: has_part CL:0000243! glial cell (sensu vertebrata)
Relationship: has_part CL:0000526! afferent neuron
Relationship: has_part CL:0000527! efferent neuron
Relationship: has_part CL:0002573! Schwann cell

ADDING CELL-TYPE RELATIONSHIPS

Once the AEO was in place, it seemed sensible to supplement the anatomical type information by annotating the leaf tissues of *simple-tissue* and the appropriate subclasses of *multi-tissue structure* with their cell-types and these are formally detailed in the cell-type ontology. This ontology is unnecessarily rich for the fairly simple annotation exercise here and a sub-ontology of the ∼100 cell-type classes that were needed (∼15% of the original ontology) was made on the basis of standard histology textbooks and incorporated within the AEO (**Figure 4**). There is no ideal relationship to convey the sense that a particular cell-type is a major but not exclusive constituent in a particular class of tissue (there may be several cell-types in such relatively simple-tissues as ganglia, epithelia, and mesenchymal domains, **Table 1**). The relationship chosen for the link was *has_part* and this carries the meaning that tissue A has some part made of cell-type B, as can be seen by inspection of the Obo file (**Table 1**). Unfortunately, browsers require that the relationships be read upward and so the link is under the cell-type rather than under the tissue (**Figure 5**).

Making the *cell-type* to *simple-tissue* relationships was usually straightforward, in the sense that one cell-type could usually be seen as the predominant type for a tissue, but this was not always so, particularly for the tissues of the nervous system such as ganglions where neurons are always accompanied by support cells.

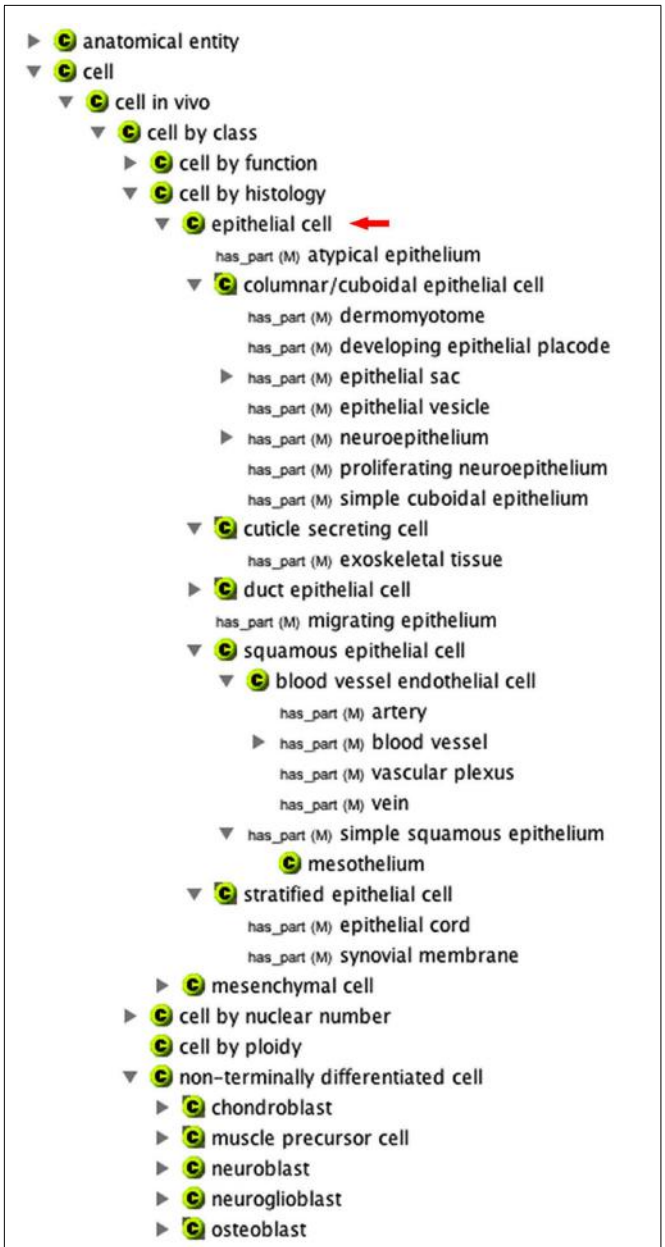


FIGURE 5 | The *cell-type* hierarchy of the AEO shown in the COBrA browser. This subset of the full cell-type ontology includes about 90 classes in all but only the epithelial cell class is expanded (red arrow). The editor uses the *has_part* relationship to show those anatomical entities (simple-tissues) which include the various epithelial cell-types.

DISCUSSION

There were two key reasons for producing the AEO: first, to provide a formal type definition for the ontology of human developmental anatomy so that it would meet modern ontology standards and, second, to enable this and perhaps other anatomy ontologies, which are mainly built from *part_of* relationships, to increase the amount of anatomical knowledge that they contain. This new ontology had, of course, to be based on the CARO scaffold, as its use is now standard for anatomy ontologies.

In practice, the only problem in using the CARO as a scaffold turned out to derive from the definition of the term *simple-tissue* (or *portion of tissue*) whose definition was "anatomical structure, that consists of similar cells and intercellular matrix, aggregated according to genetically determined spatial relationships." This definition assumes that a single structure (the anatomical term) is composed of essentially similar cells, and, while this is usually so, there are some important exceptions, particularly in the nervous system where ganglia and brain nuclei where neurons, the key functional cell-type in the nervous system: neurons in ganglia and brain nuclei are always accompanied by support cells (glia, astrocytes, etc.). Another type of structure that could, on the basis of its boundaries, be viewed as simple is the membrane bone which includes osteoblasts, osteoclasts, and osteocytes. Placing such structures within the AEO could not be done in any natural way, and the solution adopted was based on what seemed to be the most appropriate location for a user. *Membrane bone* was made a subclass of *bone*, a multi-tissue structure, while *ganglion* and *neuronal nucleus* were made subclasses of *neuronal structure*, a

simple-tissue. While these choices are not logically consistent, it is to be hoped that they will not lead to any downstream problems.

Attaching the cell-type terms to the new classes was done relatively late in the production of the AEO as it became clear that it would be quite simple and straightforward to add them from a subset of the cell-type ontology. Although there is always the concern that a user might suppose that the relationships are complete, they are not! It should be emphasized that only key cell-types have been included. Thus, for example, most tissues in an organism have associated blood vessels, nerve endings, and phagocytic cells and these have not been included. The relationship used here is *has_part* (every *endochondral bone has_part one or more osteocytes*) and this allows the cell-types to be associated with the tissues in the obo file. Browsers views this relationship in an inverted way and show the tissues associated with a cell-type (**Figure 5**).

A key part of making the ontology was using its classes to annotate the anatomical terms in the ontology of human developmental anatomy (EHDAA2) being revised from that of Hunter et al. (2003). This annotation exercise frequently demanded that new AEO terms be added and occasionally that their location be changed. As a result of annotating the EHDAA2 ontology and glancing through other vertebrate ontologies the terms seem adequate for typing vertebrate tissues. The same amount of attention has not, it should be said, been given to invertebrate ontologies and those working in this area may well find that, if they choose to use the AEO, they will need additions or changes (see below).

The only reason for producing a new ontology is that it should be useful and I hope that the integration of AEO within anatomy ontologies other than that for human developmental anatomy may prove helpful in two contexts at least. First, it would help curators who wish to annotate and users who wish to search on the basis of anatomical structure (e.g., all ducted glands). Second, it would be of value to anyone who wishes to know something about the histology of a tissue and the sort of cells that it contains. In this context, the *simple-tissue* hierarchy may be particularly useful to both groups. In addition, the AEO can rightly be seen as no more than an expansion in breadth and granularity of the CARO and it is a fair question as to whether the AEO should absorbed within the CARO with AEO ids becoming CARO ids. This is really a question that the curators of the CARO and of anatomy ontologies other than that for human developmental anatomy will need to answer; if they do decide to do this, the transfer will be easy: as there is no current overlap in id number between the CARO and AEO namespaces, it will merely require a global change in the AEO OBO file of <AEO:> to <CARO:> (provided, of course, that no new terms are added to CARO in the meantime).

The ontology is named the *Anatomical Entity Ontology* and this might seem a little ambitious, given that its focus is primarily on vertebrate and secondarily on invertebrate anatomy, with little attention so far being paid to plants and fungi anatomy. In practice, there are terms within the AEO that can be used to type such tissues (e.g., nectar and sap are non-tissue substance, cambium, and root meristem are developing tissues with stem cells, a *dictyostelium* slug *is_a* migrating tissue and hyphae *is_a* epithelial plexus). That said, the AEO does not yet contain specific plant and fungal terms and it is intended that future drafts will include appropriate type term for classifying organisms from these phyla. It is also planned that they will include semantic features automating classification (Rector, 2003; Meehan et al., 2011).

Although drafts of the AEO have been discussed with others (see Acknowledgments), the ontology will inevitably have errors and omissions. Suggestions, comments and criticisms should be sent to j.bard@ed.ac.uk. This and further versions of the ontology will be posted at and downloadable from the OBO foundry http://www.obofoundry.org/. A summary of the ontology can be found at http://www.obofoundry.org/wiki/index.php/AEO:Main_Page and this wiki will be used to post details of future changes.

ACKNOWLEDGMENTS

I thank Richard Baldock, Albert Burger, Duncan Davidson, Melissa Haendel, Gillian Morriss-Kay, and David-Osumi-Sutherland for discussion. I am grateful to Chris Mungall for always being available for email conversations, for commenting on the manuscript, regularizing the relationships, and for maintaining the obo site. I also thank the reviewers for their helpful criticisms.

REFERENCES

Aitken, S., Korf, R., Webber, B., and Bard, J. (2005). COBrA: a bio-ontology editor. *Bioinformatics* 21, 825–826.

Bard, J. B. L. (2008). "Anatomical ontologies for model animals," in *Anatomy Ontologies for Bioinformatics*, eds A. Burger, D. Davidson, and R. Baldock (Springer), 3–25.

Burger, A., Davidson, D., and Baldock, R. (eds). (2008). *Anatomy Ontologies for Bioinformatics*. New York: Springer.

Day-Richter, J., Harris, M. A., Haendel, M., The Gene Ontology OBO-Edit Working Group, and Lewis, S. (2007). Obo-Edit – an ontology editor for biologists. *Bioinformatics* 23, 2198–2200.

Haendel, M. A., Neuhaus, F., Osumi-Sutherland, D., Mabee, P., Mejino, J. L. V., Mungall, C., and Smith, B. (2008). "CARO, the common anatomy reference ontology," in *Anatomy Ontologies for Bioinformatics*, eds A. Burger, D. Davidson, and R. Baldock (New York: Springer), 327–349.

Hunter, A., Kaufman, M. H., McKay, A., Baldock, R., Simmen, M. W., and Bard, J. B. L. (2003). An ontology of human developmental anatomy. *J. Anat.* 203, 347–355.

Meehan, T. F., Masci, A. M., Abdulla, A., Cowell, L. G., Blake, J. A., Mungall, C. J., and Diehl, A. D. (2011). Logical development of the cell ontology. *BMC Bioinformatics* 12, 6. doi: 10.1186/1471-2105-12-6

Rector, A. (2003). Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proc. KCAP ACM* 2003, 121–129.

Ross, M. H., Romrell, L. J., and Kaye, G. I. (1995). *Histology: a Text and Atlas*. Baltimore: Williams & Wilkins.

Standring, S. (ed.). (2008). *Gray's Anatomy*, 40th Edn. London: Churchill Livingstone.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* 7, e1000247. doi: 10.1371/journal.pbio.1000247

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 November 2011; accepted: 28 January 2012; published online: 14 February 2012.

Citation: Bard JBL (2012) The AEO, an ontology of anatomical entities for classifying animal tissues and organs. Front. Gene. ***3****:18. doi: 10.3389/fgene.2012.00018*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 23 May 2012
doi: 10.3389/fgene.2012.00079

IMGT-ONTOLOGY 2012

Véronique Giudicelli and Marie-Paule Lefranc*

IMGT®, the international ImMunoGenetics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire, Institut de Génétique Humaine, UPR CNRS, Montpellier, France

Edited by:
John Hancock, Medical Research Council, UK

Reviewed by:
Ruth Lovering, University College London, UK
Anna Maria Masci, Duke University, USA

****Correspondence:***
Marie-Paule Lefranc, IMGT®, the international ImMunoGenetics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire, Institut de Génétique Humaine, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier cedex 5, France.
e-mail: Marie-Paule.Lefranc@igh.cnrs.fr

Immunogenetics is the science that studies the genetics of the immune system and immune responses. Owing to the complexity and diversity of the immune repertoire, immunogenetics represents one of the greatest challenges for data interpretation: a large biological expertise, a considerable effort of standardization and the elaboration of an efficient system for the management of the related knowledge were required. IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY, which represents the first ontology for the formal representation of knowledge in immunogenetics and immunoinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the seven axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope: "IDENTIFICATION," "DESCRIPTION," "CLASSIFICATION," "NUMEROTATION," "LOCALIZATION," "ORIENTATION," and "OBTENTION." The concepts of identification, description, classification, and numerotation generated from the axioms led to the elaboration of the IMGT® standards that constitute the IMGT Scientific chart: IMGT® standardized keywords (concepts of identification), IMGT® standardized labels (concepts of description), IMGT® standardized gene and allele nomenclature (concepts of classification) and IMGT unique numbering and IMGT Collier de Perles (concepts of numerotation). IMGT-ONTOLOGY has become the global reference in immunogenetics and immunoinformatics for the knowledge representation of immunoglobulins (IG) or antibodies, T cell receptors (TR), and major histocompatibility (MH) proteins of humans and other vertebrates, proteins of the immunoglobulin superfamily (IgSF) and MH superfamily (MhSF), related proteins of the immune system (RPI) of vertebrates and invertebrates, therapeutic monoclonal antibodies (mAbs), fusion proteins for immune applications (FPIA), and composite proteins for clinical applications (CPCA).

Keywords: IMGT, immunogenetics, immunoinformatics, IMGT-ONTOLOGY, immunoglobulin, antibody, T cell receptor, immune repertoire

INTRODUCTION

Immunogenetics is the science that studies the genetics of the immune system and immune responses. Among them, the adaptive immune response, acquired by vertebrates with jaws or *gnathostomata*, is characterized by an extreme diversity of the specific antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR). The potential repertoire of each individual is estimated to comprise about 2×10^{12} different IG and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity results from the complex and unique molecular synthesis and genetics of the antigen receptor chains that include DNA molecular rearrangements (combinatorial diversity) in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity) and, for the IG, somatic hypermutations (for review see Lefranc and Lefranc, 2001a,b).

Owing to the complexity and diversity of the immune repertoires and their implications in fundamental and medical research, immunogenetics represents one of the greatest challenges for data interpretation: a large biological expertise, a considerable effort of standardization and the elaboration of an efficient system for the management of the related knowledge were required. To answer that challenge, IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) was created in 1989 by the Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France (Lefranc et al., 2009; Lefranc, 2011a). IMGT® has become the global reference in immunogenetics and immunoinformatics. IMGT® is a high-quality integrated knowledge resource that provides a common access to standardized data from genome, proteome, genetics, two-dimensional (2D) and three-dimensional (3D) structures. It comprises 7 databases (sequence, gene, structure and specialist databases), 17 online tools and more than 15,000 pages of web resources (Lefranc et al., 2009).

IMGT® has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY started in 1989 and, since then, in constant evolution and extension (Giudicelli and Lefranc, 1999; Lefranc et al., 2004, 2005a, 2008; Duroux et al., 2008; Lefranc, 2011b,c,d,e,f, 2013). IMGT-ONTOLOGY represents the first ontology for the formal representation of knowledge in

immunogenetics and immunoinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the seven axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope: "IDENTIFICATION," "DESCRIPTION," "CLASSIFICATION," "NUMEROTATION," "LOCALIZATION," "ORIENTATION," and "OBTENTION" (Duroux et al., 2008). These axioms postulate that any object, any process and any relation has to be identified, described, classified, numbered, localized, and orientated, and that the way it is obtained can be characterized. The IMGT-ONTOLOGY concepts were generated from these axioms. The concepts of identification, description, classification, and numerotation led to the elaboration of the IMGT® standards that constitute the IMGT Scientific chart: IMGT® standardized keywords (concepts of identification), IMGT® standardized labels (concepts of description), IMGT® standardized IG and TR gene and allele nomenclature (concepts of classification) and IMGT unique numbering and IMGT Collier-de-Perles (concepts of numerotation). One major feature of IMGT-ONTOLOGY is the formalization of the specific relations that link, on a semantic point of view, the different concepts and capture the immunogenetics complexity. These relations are fundamental for data consistency and biological interpretation.

MATERIALS AND METHODS

An ontology is defined as "an explicit specification of a conceptualization" (Gruber, 1993; Guarino and Giaretta, 1995; Guarino, 1997). The building of IMGT-ONTOLOGY has consisted in the conceptualization and in the formalization of the related knowledge in immunogenetics, and in the definition of the relations between concepts. The first concepts were defined as "relevant and fundamental criteria which are needed to characterize IG and TR data" (Giudicelli and Lefranc, 1999). Since then, the IMGT-ONTOLOGY concepts have been largely extended to molecular components other than IG and TR, that include major histocompatibility (MH) proteins of humans and other vertebrates, proteins of the immunoglobulin superfamily (IgSF), and MH superfamily (MhSF), related proteins of the immune system (RPI) of vertebrates and invertebrates, therapeutic monoclonal antibodies (mAbs), fusion proteins for immune applications (FPIA), and composite proteins for clinical applications (CPCA).

Concepts are characterized by their properties which may be simple attributes or relations between concepts. The relation of subsumption (is_a) allows to structure the IMGT-ONTOLOGY concepts, and to represent them as nodes of the graph with their level of granularity. The concepts that correspond to the finest level of granularity (and the highest level of precision) in branches of the graph are designated as "leafconcept." Concepts from which a hierarchy is generated with several levels before reaching the leafconcepts are designated as "highconcept."

IMGT-ONTOLOGY is being formalized in OWL-DL[1] language using the Protégé editor[2] (Noy et al., 2003). The formalized concepts of identification are available for downloading or browsing on the National Center for Biomedical Ontology (NCBO) BioPortal[3] (Noy et al., 2009; Musen et al., 2012) and on the IMGT® web site (http://www.imgt.org; Lefranc, 2011a,b,c,d,e,f).

The semantic relations (other than subsumption) are formalized as OWL object properties (see OWL 2 Web Ontology Language http://www.w3.org/TR/owl-primer/): Object properties allow to link specifically two concepts through the statement "Subject > Property > Object" where "Subject" is the concept being characterized by the object property, "Property" the name of a given property defined in the ontology and "Object" the name of the concept that is linked. These properties are restricted using in particular universal quantification (all connected individuals by the property must be instances of a given class), existential quantification (all individuals of the class for which the property is defined are connected to at least one individuals of the class mentioned in the restriction) and cardinality restrictions (quantification of the number of connected individual with the property). These relations can be displayed on NCBO BioPortal in "IMGT-ONTOLOGY > Terms > Details" page. They are indicated in the "Equivalent Class" section if they are necessary and sufficient to define the concept, or in the "Sub Class Of" section if they are necessary only (for instance, the relations "is_defined_by" and "_has_" of the "D-gene" (which is a "Molecule_EntityType" leafconcept, see below "Molecule_EntityType" Concept), are examples of relations in "Equivalent Class" and "Sub Class Of" sections, respectively). The formalization of these relations highlights and focuses on the dependencies between the terms that are closely interconnected at the level of immunogenetics knowledge and set up the constraints that must be respected in the IMGT® databases and tools and in immunoinformatics.

RESULTS

IMGT-ONTOLOGY IDENTIFICATION AXIOM

The IDENTIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al., 2008) postulates that, for molecular components, any molecule and its relations have to be identified (Lefranc, 2011b). IMGT-ONTOLOGY concepts of identification generated from the IDENTIFICATION axiom led to the IMGT® standardized keywords for molecular components (IG, TR, MH, RPI, FPIA, and CPCA) in IMGT® databases and tools.

IMGT-ONTOLOGY concepts of identification

"Molecule_EntityType" concept. The objective of IMGT-ONTOLOGY was to identify the type of any molecular entity at each step of its synthesis. An insight of the knowledge related to the synthesis of an IG is schematized in **Figure 1**. It illustrates the concept of "Molecule_EntityType" and the other related concepts of identification and the relations that link them.

The "Molecule_EntityType" concept is fully defined by the concepts of "MoleculeType," "GeneType," and "ConfigurationType" (**Figure 2**).

– "MoleculeType" allows one to identify the type of molecule, based on the type of the constitutive elements and on the concepts of obtention. The "MoleculeType" concept comprises four major leafconcepts: "gDNA," "mRNA," "cDNA," "protein."

[1]http://www.w3.org/TR/owl2-overview/
[2]http://protege.stanford.edu
[3]http://bioportal.bioontology.org/ontologies/1491

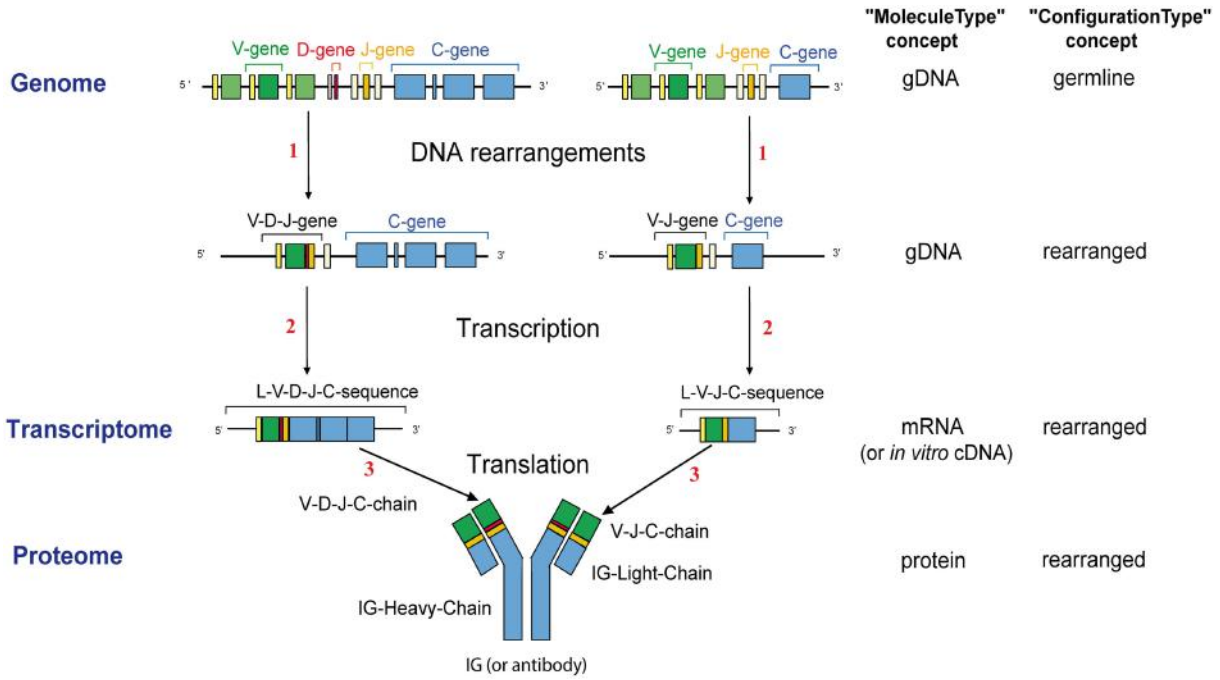


FIGURE 1 | An example of knowledge at the molecular level: the synthesis of an IG or antibody in humans, described in (Lefranc, 2011a). "gDNA," "mRNA," and "protein" are types of molecules ("MoleculeType") that are involved in the IG or TR synthesis, "germline" and "rearranged" are types of configuration ("ConfigurationType") [the configuration of C-gene is "undefined" (not shown)]. A molecule entity type characterizes a unique conformation of a molecular component at each step of its biosynthesis, which is defined by a type of molecule, a type of configuration and type(s) of genes. The 10 leafconcepts of "Molecule_EntityType" identified during the IG synthesis (e.g., V-gene, V-D-J-gene, L-V-D-J-C-sequence) are shown. Main steps of the antigen receptor synthesis are indicated with numbers. (1) DNA rearrangements (is_rearranged_into), (2) Transcription (is_transcribed_into), (3) Translation (is_translated_into) (IMGT Repertoire, http://www.imgt.org).

– "GeneType" allows one to identify the type of gene. The "GeneType" concept comprises six leafconcepts: "variable" (V), "diversity" (D), "joining" (J), and "constant" (C) are the four gene types specific of IG and TR, and "conventional-with-leader," "conventional-without-leader" are the two gene types of conventional genes.
– "ConfigurationType" allows one to identify the type of configuration of a gene, and by extension, the type of configuration of the Molecule_EntityType leafconcepts that contain it. The "ConfigurationType" concept comprises four leafconcepts: "germline," "partially-rearranged" and "rearranged" for V, D, and J genes and the molecule entities that contain them, and "undefined" for C and conventional genes and related entities.

The "Molecule_EntityType" concept comprises 38 leafconcepts (**Table 1**). For examples, "V-gene" identifies, for gDNA, molecule entities with a germline V gene, "V-D-J-gene" identifies, for gDNA, molecule entities with rearranged V, D, and J genes, and "L-V-D-J-C-sequence" identifies, for cDNA, molecule entities with rearranged V, D, and J genes spliced to a C gene. The four "MoleculeUnit" leafconcepts that are "gene" (10), "transcript" (11), "sequence" (11), and "chain" (6) identify the type of entities based on the "MoleculeType" only, as indicated by the suffix (**Table 1**).

In addition to the relation "is_defined_by," a "Molecule_EntityType" "has" properties identified in the "FunctionalityType" and "StructureType" concepts (**Figure 2**).

– "FunctionalityType" is a concept of identification that allows one to identify, whatever the molecule type (gDNA, cDNA, mRNA, or protein), the type of functionality of a Molecule_EntityType leafconcept. The "FunctionalityType" concept comprises five leafconcepts: "functional," "ORF" (open reading frame) and "pseudogene" identify the functionality of Molecule_EntityType leafconcepts in undefined configuration (conventional genes and IG and TR C genes), or in germline configuration (IG and TR V, D, and J genes before DNA rearrangements); "productive" and "unproductive" identify the functionality of Molecule_EntityType leafconcepts in rearranged or partially-rearranged configuration (IG and TR entities after DNA rearrangements, and by extension fusion entities resulting from translocations, and hybrid entities obtained by biotechnology molecular engineering).
– "StructureType" is a concept of identification that allows one to identify, whatever the molecule type (gDNA, cDNA, mRNA, protein), the type of structure of Molecule_EntityType leafconcepts.

The semantic relations of "Molecule_EntityType" are formalized as properties (in OWL).

"ChainType," "DomainType," and "ReceptorType" concepts. One of the goals of IMGT-ONTOLOGY has been to represent knowledge in order to manage molecular components from

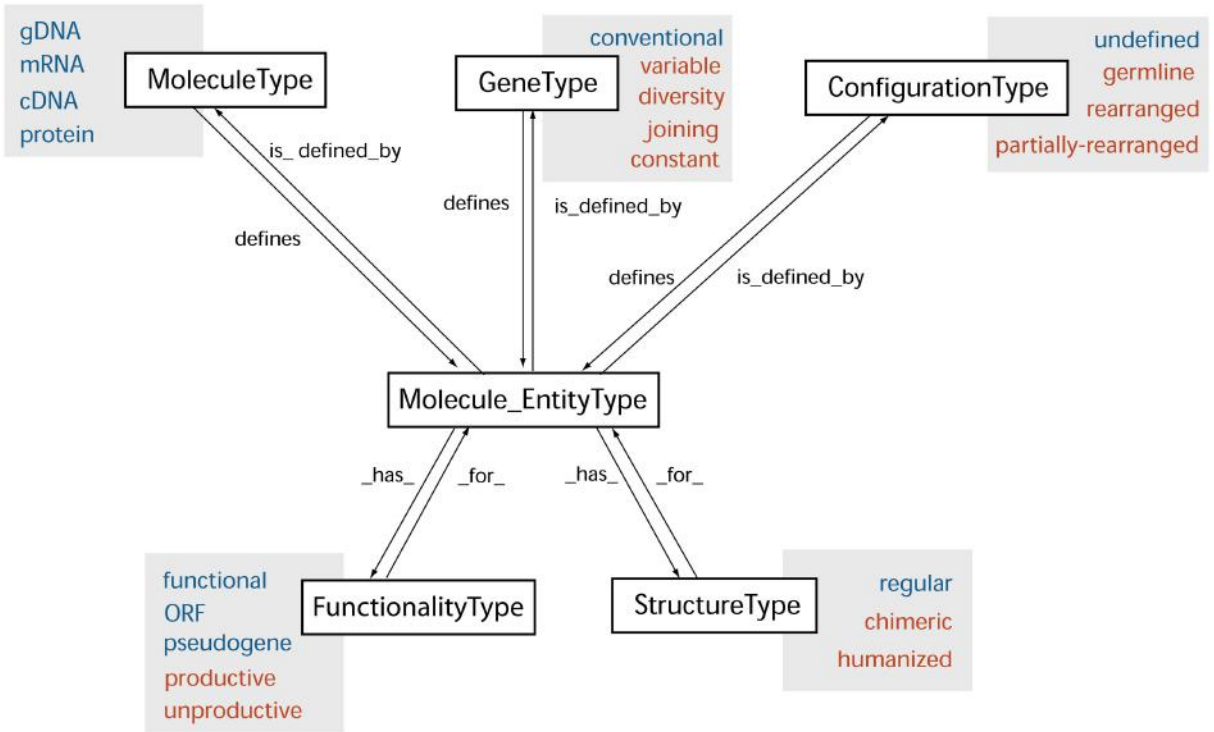


FIGURE 2 | The "Molecule_EntityType" concept with its relations. The "Molecule_EntityType" concept is defined by the "MoleculeType," "GeneType," and "ConfigurationType" concepts of identification and has properties identified in the "FunctionalityType" and "StructureType" concepts (IDENTIFICATION axiom). Arrows indicate reciprocal relations "is_defined_by" and "defines," "_has_," and "_for_." Leafconcepts are general (online in blue) or specific of the IG and TR (online in red). The "Molecule_EntityType" concept has 38 leafconcepts (or keywords in the IMGT® databases and tools). Only a few examples of the "StructureType" leafconcepts (or keywords in the IMGT® databases and tools) are shown (see details in Lefranc, 2011b).

sequences to 3D structures in IMGT® databases and tools. The three concepts "ChainType," "DomainType," and "ReceptorType" have been fundamental in that knowledge representation.

"ChainType" is a concept of identification that allows one to identify the type of chain. "ChainType" is a "highconcept" that comprises four levels (**Figure 3**): "MolecularComponentLevelChainType," "ReceptorLevelChainType," "ClassLevelChainType," and "GeneLevelChainType." The concepts are organized in an acyclic graph based on the subsumption relation, the depth of which depends on the precision that needs to (or that can be) reported for the data identification. The finest level of granularity, the "GeneLevelChainType" concept, identifies the type of chain by reference to the gene(s) which code(s) the chain. It represents the main concept for a very precise identification because it establishes a relationship with "Gene" (concept of classification) (the reciprocal relations are: "is_coded_by" and "codes"). The number of "ChainType" leafconcepts of the "GeneLevelChainType" depends on the number of functional genes and ORF ("FunctionalityType") per haploid genome, in a given species (in the case of the IG and TR genes, it is the number of functional and ORF C genes which is taken into account).

The "ChainType" concept is defined by the "Molecule_EntityType" and the "DomainType" concepts of identification, and also defined by concepts of classification (see IMGT-ONTOLOGY CLASSIFICATION Axiom) as the type of chain depends on the taxon (**Figure 4**). "DomainType" allows one to identify the type of domain. A domain is a chain subunit characterized by its three-dimensional (3D) structure, and by extension its amino acid sequence and the nucleotide sequence which encodes it.

The "ChainType" concept represents a key concept that allows to link the "Molecule_EntityType" (sequences in databases) to the concept of "ReceptorType" (3D structures in databases; **Figure 4**). "ReceptorType" allows one to identify the type of receptor. "ReceptorType" is defined by the "ChainType" leafconcept(s) that identify the associated chains of a receptor. "ReceptorType" is a "highconcept" with a hierarchy of four levels of granularity (depending on the "ChainType" hierarchy). The "ReceptorType" concept has properties identified in the "FormatType," "SpecificityType," and "FunctionType" concepts (**Figure 4**; Lefranc, 2011b).

IMGT® standardized keywords in databases and tools

The leafconcepts of identification are IMGT® standardized keywords in the IMGT® databases and tools (Lefranc, 2005). The list of IMGT® standardized keywords is available from the IMGT/LIGM-DB database (Giudicelli et al., 2006) query page (IMGT® Home page; http://www.imgt.org) and in the IMGT Scientific chart at http://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGT3Dkeywords.html. More than 325 IMGT® standardized keywords (189 for sequences and 137 for

Table 1 | "Molecule_EntityType" leafconcepts and related concepts. The 38 "Molecule_EntityType" leafconcepts are shown with the leafconcepts of "GeneType," "ConfigurationType," and "MoleculeType" that define them. The four leafconcepts of "MoleculeUnit" are based on "MoleculeType" only.

MoleculeUnit leafconcepts	Molecule_EntityType leafconcepts	GeneType leafconcepts	ConfigurationType leafconcepts	MoleculeType leafconcepts
gene	V-gene*	V	germline	gDNA
	D-gene*	D		
	J-gene*	J		
	J-C-gene	J, C		
	C-gene*	C	undefined	
	conventional-gene	conventional		
	V-D-gene	V, D	partially-rearranged	
	D-J-gene	D, J		
	V-J-gene*	V, J	rearranged	
	V-D-J-gene*	V, D, J		
transcript	L-V-transcript	V	germline	mRNA
	D-transcript	D		
	J-transcript	J		
	J-C-transcript	J, C		
	C-transcript	C	undefined	
	L-nt-transcript	conventional		
	nt-transcript	conventional		
	L-V-D-transcript	V, D	partially-rearranged	
	D-J-C-transcript	D, J, C		
	L-V-J-C-transcript	V, J, C	rearranged	
	L-V-D-J-C-transcript	V, D, J, C		
sequence	L-V-sequence	V	germline	cDNA
	D-sequence	D		
	J-sequence	J		
	J-C-sequence	J, C		
	C-sequence	C	undefined	
	L-nt-sequence	conventional		
	nt-sequence	conventional		
	L-V-D-sequence	V, D	partially-rearranged	
	D-J-C-sequence	D, J, C		
	L-V-J-C-sequence*	V, J, C	rearranged	
	L-V-D-J-C-sequence*	V, D, J, C		
chain	L-AA-chain	conventional	undefined	Protein
	AA-chain	conventional		
	L-V-J-C-chain	V, J, C	rearranged	
	L-V-D-J-C-chain	V, D, J, C		
	V-J-C-chain*	V, J, C		
	V-D-J-C-chain*	V, D, J, C		

Indicates the 10 leafconcepts that are the most classical representatives of IG and TR identification. These leafconcepts are illustrated in **Figure 1.*

3D structures) were precisely defined. They represent the controlled vocabulary assigned during the annotation process and allow standardized search criteria for querying the IMGT® databases and for the extraction of sequences and 3D structures. IMGT/HighV-QUEST, the IMGT® tool for analysis of IG and TR nucleotide sequences obtained from next generation sequencing (NGS; Alamyar et al., 2012), provides an evaluation of the configuration ("ConfigurationType") and, accordingly, of the sequence functionality ("FunctionalityType"): such precision and standardization in the NGS results are of the utmost importance for the reuse of data for the statistical analyses required for the comparison of immune repertoires (Prabakaran et al., 2012) and for data interpretation.

IMGT-ONTOLOGY DESCRIPTION AXIOM

The DESCRIPTION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al., 2008) postulates that, for molecular components, any molecule and its relations have to be described (Lefranc, 2011c). IMGT-ONTOLOGY concepts of description generated from the DESCRIPTION axiom led

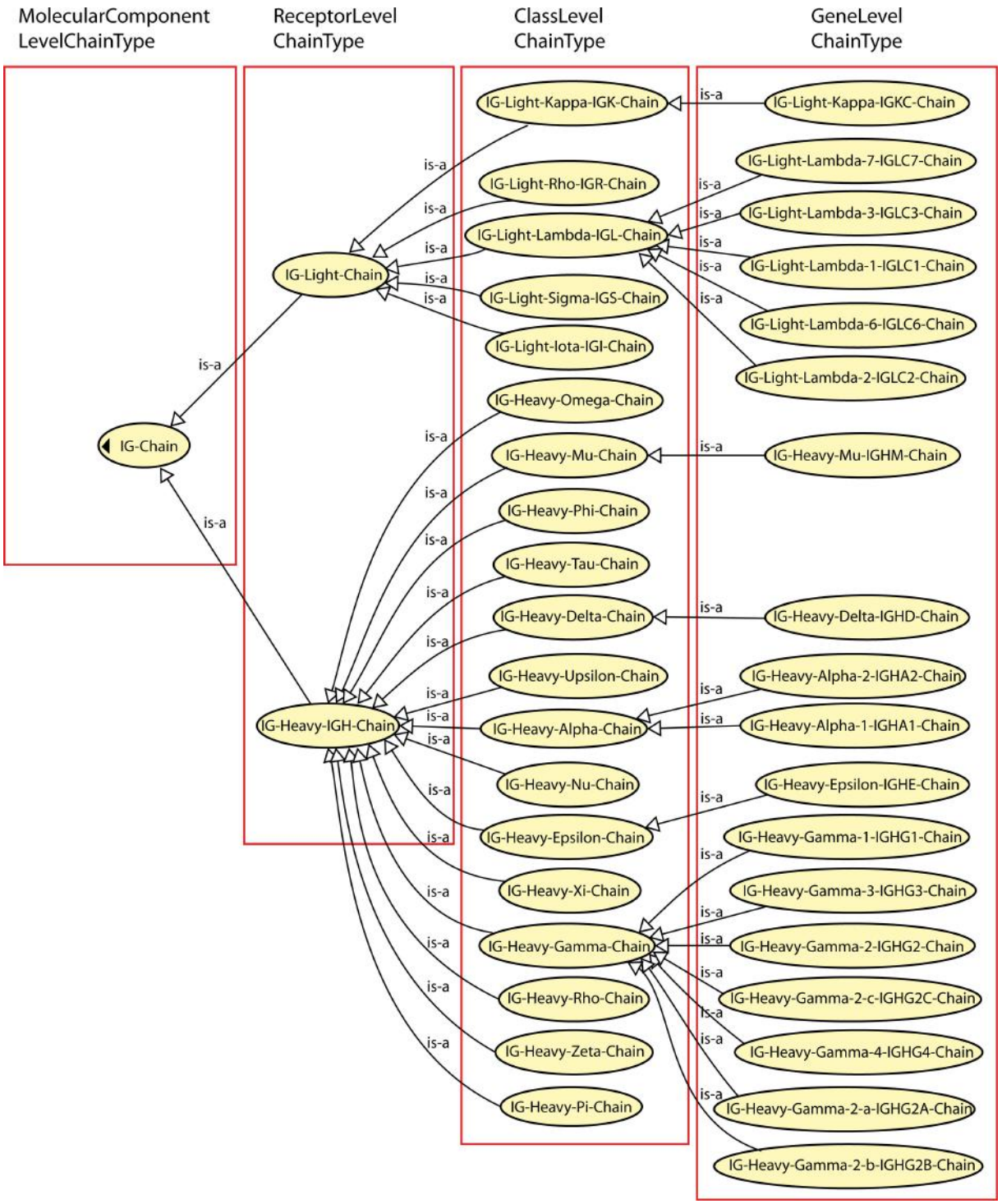


FIGURE 3 | "ChainType" "highconcept." The hierarchy of "ChainType" for the identification of IG chains comprises four levels of granularity which are associated with an increasing level of precision: the "MolecularComponentLevelChainType," the "ReceptorLevelChainType," the "ClassLevelChainType," and the "GeneLevelChainType."

to the IMGT® standardized labels for molecular components (IG, TR, MH, RPI, FPIA, and CPCA) in IMGT® databases and tools.

IMGT-ONTOLOGY concepts of description

Concepts of description have been progressively elaborated in order to take into account the entities of the different steps of

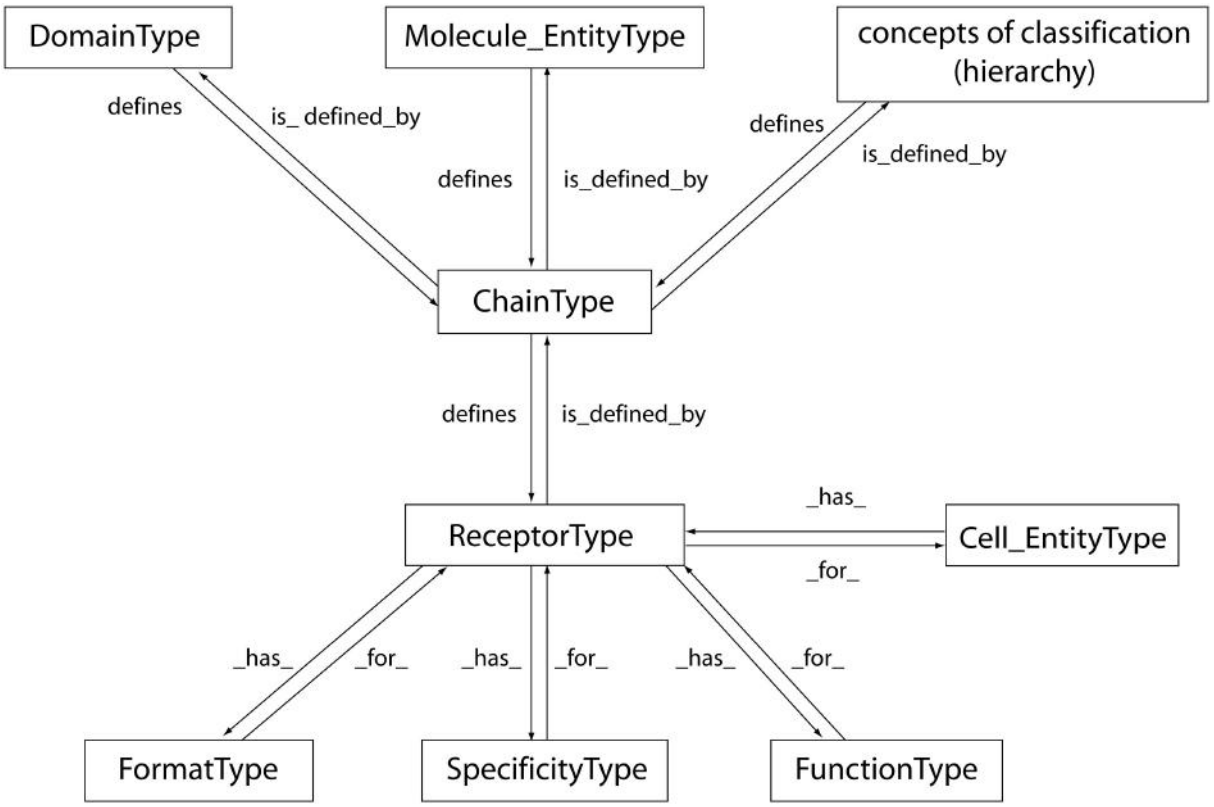


FIGURE 4 | The "ReceptorType" concept with its relations. The "ReceptorType" concept is defined by the "ChainType" concept of identification and has properties identified in the "FormatType", "SpecificityType," and "FunctionType" concepts (IDENTIFICATION axiom). The "ChainType" concept is itself defined by the "Molecule_EntityType" and "DomainType" concepts and by concepts of classification organized in a hierarchy (see CLASSIFICATION axiom). Arrows indicate reciprocal relations "is_defined_by" and "defines," "_has_," and "_for_" (see details in Lefranc, 2011b). The "ChainType" and "ReceptorType" concepts have different levels of granularity (up to four) and are highconcepts. The reciprocical relations between "ReceptorType" and "Cell_EntityType" concepts are "_has_" and "_for_." The "Cell_EntityType" (not developped in the current version) is part of the "CellularComponent" concept (Pappalardo et al., 2010).

the molecular synthesis of the antigen receptors (IG and TR) and, more generally, of all molecular components and to describe all motifs of biological interest of sequences and 2D and 3D structures in databases and tools.

"Molecule_EntityPrototype" concept. The "Molecule_EntityPro totype" is a concept, generated from the DESCRIPTION axiom, that provides the description of the "Molecule_EntityType" concept (IDENTIFICATION axiom). There are as many leaf-concepts in the "Molecule_EntityPrototype" as there are leaf-concepts in the "Molecule_EntityType." Thus the "Molecule_EntityPrototype" comprises 38 leafconcepts that describe the organization of each entity with its constitutive motifs and relations. Each "Molecule_EntityPrototype" leafconcept is linked to a "Molecule_EntityType" leafconcept by the reciprocal relations "describes" and "is_described_by." For example, a "V-gene" is described by "V-GENE," and a "V-D-J-gene" by "V-D-J-GENE." Leafconcepts of description (labels in the IMGT® databases and tools) are written in capital letters.

Prototypes and relations between concepts of description. In order to visualize the organization of each entity, prototypes were defined. A prototype is a graphical representation of a "Molecule_EntityPrototype" leafconcept. Two prototypes of "V-GENE" and "V-D-J-GENE" are shown in **Figure 5** as examples of a germline entity and of a rearranged entity, respectively. Twenty-seven labels for "V-GENE" and 33 labels for "V-D-J-GENE" (20 of them being shared by the two prototypes), on a total of 277 different labels for sequences in IMGT/LIGM-DB, are necessary and sufficient for a complete description of these prototypes. The organization of a prototype is based on the relations that order two labels.

IMGT-ONTOLOGY formalizes the topological relations that define the relative position of two labels. A set of twelve relations are necessary and sufficient to describe the relations between labels in a prototype (Duroux et al., 2008; Lane et al., 2010; **Table 2**). The reciprocal relations "is_in_5_prime_of" and "is_in_3_prime_of" describe the relative position of labels on a 5′–3′ DNA strand when there is no intersection or contiguity between labels (Lane et al., 2010).

IMGT® standardized labels in databases and tools

The leafconcepts of description are IMGT® standardized labels in the databases and tools (Lefranc, 2005). The IMGT®

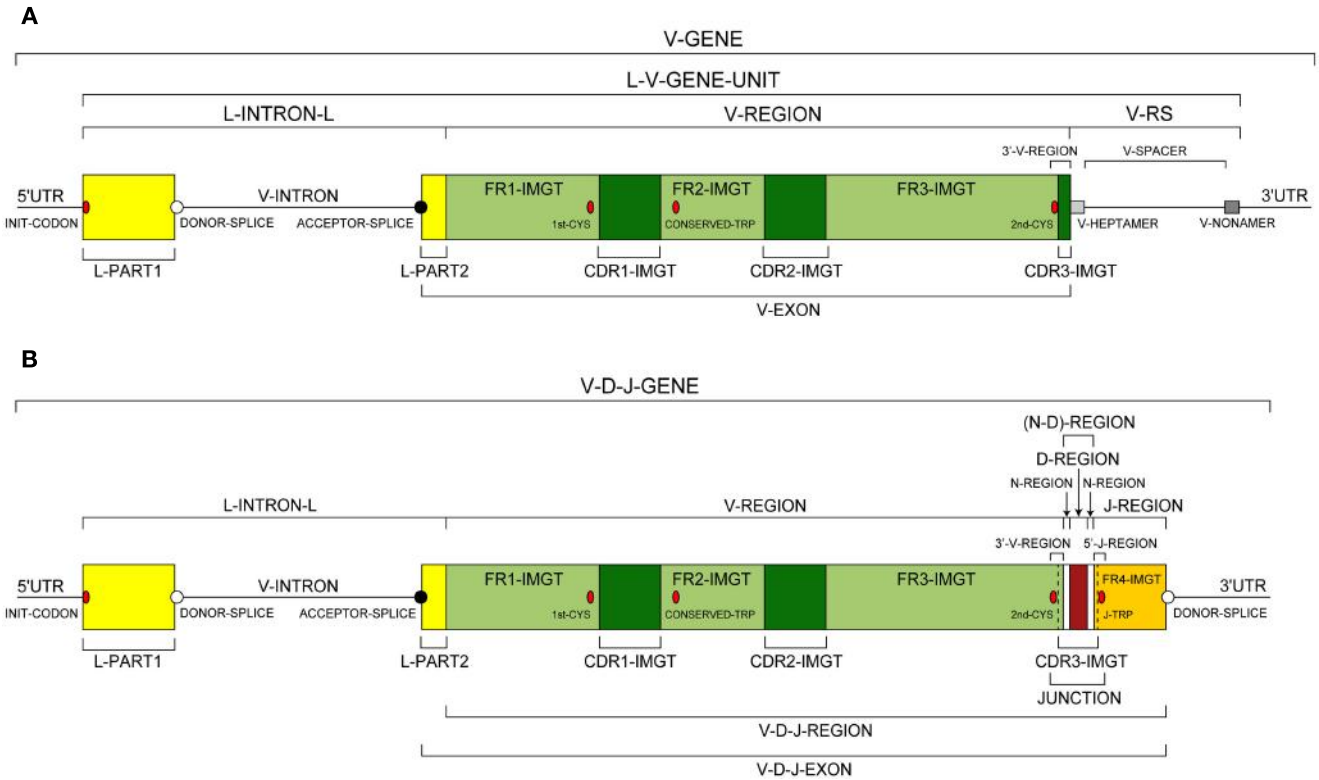


FIGURE 5 | Prototype or graphical representation of two "Molecule_EntityPrototype" leafconcepts. (A) "V-GENE." **(B)** "V-D-J-GENE." Thirty-nine labels (27 for for "V-GENE" and 33 for "V-D-J-GENE" of which 20 are shared) and 12 relations are necessary and sufficient for a complete description of these prototypes (Lefranc, 2011c).

Table 2 | IMGT-ONTOLOGY relations between labels used for the description of prototypes.

Relation	Reciprocal relation
"adjacent_at_its_5_prime_to"	"adjacent_at_its_3_prime_to"
"included_with_same_5_prime_in"	"includes_with_same_5_prime"
"included_with_same_3_prime_in"	"includes_with_same_3_prime"
"overlaps_at_its_5_prime_with"	"overlaps_at_its_3_prime_with"
"included_in"	"includes"
"is_in_5_prime_of"	"is_in_3_prime_of"

standardized labels are available from the IMGT/LIGM-DB database (Giudicelli et al., 2006) query page (IMGT® Home page; http://www.imgt.org) and in the IMGT Scientific chart at: http://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGT3Dkeywords.html (definitions of these labels are available at: http://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGT3Dlabeldef.html). More than 560 IMGT® standardized labels (277 for sequences and 285 for 3D structures) were precisely defined.

IMGT/Automat, the IMGT® tool for the annotation of rearranged cDNA (Giudicelli et al., 2005a) implements corresponding labels and prototypes. IMGT® standardized labels and the organization of "Molecule_EntityPrototype" have recently been implemented in IMGT/LIGMotif for the automation of the annotation of large genomic sequences (Lane et al., 2010). A set of specific labels was defined to describe the different organizations of IG and TR genes in clusters at the scale of the locus or of the chromosome.

IMGT-ONTOLOGY CLASSIFICATION AXIOM

The CLASSIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al., 2008) postulates that, for molecular components, any molecule and its relations have to be classified (Lefranc, 2011d). IMGT-ONTOLOGY concepts of classification generated from the CLASSIFICATION axiom led to the IMGT® standardized IG and TR gene and allele nomenclature.

IMGT-ONTOLOGY concepts of classification

The IMGT® standardized gene and allele nomenclature is based on the concepts of classification, generated from the CLASSIFICATION axiom, which defines the principles for the nomenclature of highly polymorphic multigene loci and families. In particular, the concepts of classification have allowed to classify the genes whatever the antigen receptor (IG or TR), whatever the locus (e.g., for mammals, immunoglobulin heavy IGH, immunoglobulin kappa IGK, immunoglobulin lambda IGL, T cell receptor alpha TRA, T cell receptor beta TRB, T cell receptor gamma TRG, and T cell receptor delta TRD), whatever the gene configuration (germline,

undefined, or rearranged), and whatever the species, from fish to human. Among the concepts of classification, the "Group," "Subgroup," "Gene," and "Allele" concepts are essential for the IMGT® gene nomenclature (Giudicelli and Lefranc, 1999). They are shown with their semantic relations in **Figure 6** that are used for the V gene designation.

IMGT® standardized IG and TR gene and allele nomenclature

In the context of the gene and allele classification, ontological principles defined in IMGT-ONTOLOGY have preceded the IMGT® standardized gene and allele nomenclature. This has been true for the human genes, and all IMGT® IG and TR gene names (Lefranc, 2000a,b; Lefranc and Lefranc, 2001a,b) were defined before the complete human genome sequencing (Lander et al., 2001; Venter et al., 2001). This is still the case for newly sequenced genomes and the denomination of IG and TR genes from a newly sequenced species is considerably facilitated by the preexisting nomenclature principles and rules. Full IMGT® standardized gene name comprises the latin names of the genus and species (e.g., *Homo sapiens* IGHV1-2). Gene names used in natural language and in publications may include abbreviation if needed for tables or figures (6-letter code for genus and species, 9-letter code for genus, species, and subspecies).

Interoperability between IMGT, HGNC, and NCBI

Since the creation of IMGT®, the international ImMunoGeneTics information system® in 1989, at New Haven during the 10th Human Genome Mapping Workshop (HGM10), the standardized classification and nomenclature of the IG and TR of humans and other vertebrate species have been under the responsibility of the IMGT Nomenclature Committee (IMGT-NC). The IMGT® gene nomenclature for human IG and TR genes (Lefranc, 2000a,b; Lefranc and Lefranc, 2001a,b) was approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 (Wain et al., 2002) and endorsed by the World Health Organization-International Union of Immunological Societies (WHO-IUIS; Lefranc, 2007, 2008). IMGT® IG and TR gene names are the official international reference and have been entered in IMGT/GENE-DB, the IMGT® gene database (Giudicelli et al., 2005b), in the Human Genome Database (GDB; Letovsky et al., 1998), in LocusLink at the National Center for Biotechnology Information (NCBI) in 1999–2000 (Maglott et al., 2000), in NCBI Entrez Gene when this gene database superseded LocusLink (Maglott et al., 2007), in NCBI Gene and in NCBI MapViewer, in Ensembl at the European Bioinformatics Institute (EBI) in 2006 (Hubbard et al., 2002), and in the Vega Genome Browser at the Wellcome Trust Sanger Institute (Ashurst et al., 2005). Amino acid sequences of human IG and TR C genes were provided to UniProt in 2008 (Bairoch et al., 2009). Close collaborations have been developed to maintain interoperability between the databases, with HGNC (Wain et al., 2004; Bruford et al., 2008), NCBI Gene (Maglott et al., 2011), Ensembl, Vega (Wilming et al., 2008), the Mouse Genomic Nomenclature Committee (MGNC), the Nomenclature Committees of newly sequenced genomes, for example, ZFIN for the zebrafish *Danio rerio* (Bradford et al., 2011) or external team contribution, for example, TRB locus of the rhesus macaque *Macaca mulatta* (Greenaway et al., 2009). IG and TR genes are also integrated in the HUGO ontology and NCI Metathesaurus available on the NCBO BioPortal[4]. Mapping between the HUGO ontology and IMGT-ONTOLOGY will be developed with the formalization of the concepts of classification in OWL.

IMGT-ONTOLOGY NUMEROTATION AXIOM

The NUMEROTATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al., 2008) postulates that, for molecular components, any molecule and its relations have to be numbered (Lefranc, 2011e,f). Two major IMGT-ONTOLOGY concepts of numerotation generated from the NUMEROTATION

[4]http://bioportal.bioontology.org/ontologies/

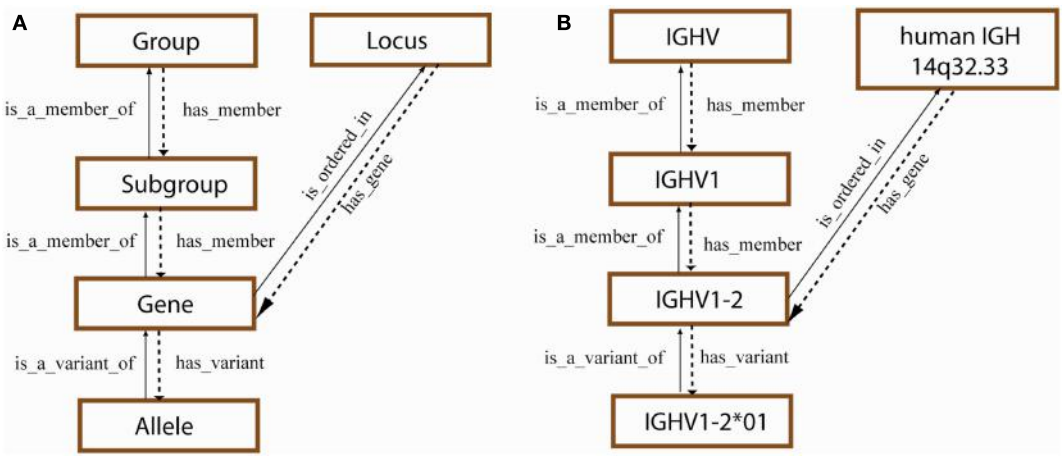


FIGURE 6 | Concepts of classification for gene and allele nomenclature (generated from the IMGT-ONTOLOGY CLASSIFICATION axiom) (Duroux et al., 2008; Lefranc, 2011d). (A) Hierarchy of the concepts of classification and their relations (Giudicelli and Lefranc, 1999). The "Locus" concept is a concept of localization (LOCALIZATION axiom). **(B)** Example of leafconcepts for each concept of classification. They are associated with a "TaxonRank" level, and more precisely for the "Gene" and "Allele" concepts with a leafconcept of "Species" (here, *Homo sapiens*).

axiom comprises the "IMGT_unique_numbering" and "IMGT_Collier_de_Perles" (IMGT unique numbering and IMGT Colliers de Perles in IMGT® databases and tools).

"IMGT_unique_numbering"

The "IMGT_unique_numbering" concept (Lefranc, 2011e) defines a systematic and coherent numbering (amino acids and codons) for the description of "DomainType" leafconcepts. The "IMGT_unique_numbering" was originally defined for the IG and TR V-DOMAIN (Lefranc, 1997). It provides a standardized delimitation of the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT), and therefore allows to correlate each position (amino acid or codon) with the structure (beta strand, loop, beta turn) and the function (antigen binding) of the V-DOMAIN. FR-IMGT and CDR-IMGT lengths became a major property of the IG and TR V-DOMAIN. The "IMGT_unique_numbering" concept has been extended to the V-LIKE-DOMAIN of the IgSF other than IG and TR (Lefranc, 1999; Lefranc et al., 2003), to the C domain (C-DOMAIN of IG and TR and C-LIKE-DOMAIN of IgSF other than IG and TR; Lefranc et al., 2005b) and to the G domain (G-DOMAIN of MH and G-LIKE-DOMAIN of MhSF other than MH) (Lefranc et al., 2005c). Thus, the "IMGT_unique_numbering" concept allows to number domain types that are characteristic of protein superfamilies, whatever the species, the molecule type or the chain type. Three leafconcepts have been defined for the variable (V) domain, the constant (C) domain, and the groove (G) domain: "IMGT_unique_numbering_for_V_domain" (Lefranc, 1997, 1999; Lefranc et al., 2003) and "IMGT_unique_numbering_for_C_domain" (Lefranc et al., 2005b) of the IG, TR and IgSF, and "IMGT_unique_numbering_for_G_domain" (Lefranc et al., 2005c) of the MH and MhSF.

"IMGT_Collier_de_Perles"

The "IMGT_Collier_de_Perles" concept (Lefranc, 2011f) corresponds to the graphical 2D representation of domains based on the set of rules defined by the "IMGT_unique_numbering." This original and unique approach allows one to bridge the gap between sequences and 2D and 3D structures and greatly facilitates the domain comparison, position per position. Three leafconcepts are defined: "IMGT_Collier_de_Perles for_V_domain" (Lefranc, 1999; Lefranc et al., 2003), "IMGT_Collier_de_Perles_for_C_domain" (Lefranc et al., 2005b) and "IMGT_Collier_de_Perles for_G_domain" (Lefranc et al., 2005c).

Figure 7 shows graphical representations of "IMGT_Collier_de_Perles_for_V_domain" (Lefranc et al., 2003). The five highly conserved amino acids found in IG and TR V domains, whatever the species and molecule type, are highlighted (online in red letters): at position 23 (1st-CYS, or first conserved cysteine C), 41 (CONSERVED-TRP, or conserved tryptophan W), 89 (hydrophobic amino acid, here methionine M), 104 (2nd-CYS, or second conserved cysteine C), and 118 (here J-PHE, or J-REGION tryptophan W). This leafconcept allows, for the first time, to compare domains of IG and TR (V-DOMAIN) and of IgSF proteins other than IG or TR (V-LIKE-DOMAIN), on one layer (facilitating comparison with sequences) or on two layers (bridging comparison with 3D structures).

Figure 8 shows graphical representations of "IMGT_Collier_de_Perles_for_G_domain" (Lefranc et al., 2005c). This leafconcept allows, for the first time, to compare domains of the same chain (G-ALPHA1 and G-ALPHA2 of MH1), domains of different chains of the same receptor (G-ALPHA and G-BETA of MH2), or domains of MhSF proteins other than MH (G-ALPHA1-LIKE and G-ALPHA2-LIKE of RPI-MH1Like).

IMGT unique numbering and IMGT Collier de Perles in databases and tools

The IMGT unique numbering and the IMGT Colliers de Perles are used for the numbering of both the codons (in nucleotide sequences) and the amino acids (in protein sequences and structures; Ruiz and Lefranc, 2002; Garapati and Lefranc, 2007; Kaas and Lefranc, 2007; Kaas et al., 2007). By facilitating the comparison of residues between sequences, the IMGT unique numbering and the IMGT Colliers de Perles have been the basis for the description of the IG and TR gene allelic polymorphism and for the studies of IG somatic hypermutations in V-DOMAIN. They represent a major breakthrough for the analysis and the comparison of the huge repertoires of antigen receptors (potentially 2×10^{12} per individual). Indeed, the IMGT unique numbering and the IMGT Colliers de Perles represent a key component in immunogenetics studies by creating a strong and reliable interoperability between the IMGT® databases, tools, and web resources (Lefranc et al., 2009).

Rules for the IMGT unique numbering are implemented in IMGT® online tools: for the analysis of IG and TR rearranged cDNA sequences by IMGT/V-QUEST (Brochet et al., 2008; Giudicelli et al., 2011) and IMGT/JunctionAnalysis (Yousfi Monod et al., 2004; Bleakley et al., 2006; Giudicelli and Lefranc, 2011), for the analysis of cDNA sequences from high-throughput NGS sequencing by IMGT/HighV-QUEST (Alamyar et al., 2012) and for the analysis of amino acid sequences and 2D structures by IMGT/DomainGapAlign (Ehrenmann and Lefranc, 2011a), IMGT/DomainDisplay and IMGT/Collier-de-Perles (Ehrenmann et al., 2011). They are also implemented in IMGT® databases, and particularly in IMGT/3Dstructure-DB (Ehrenmann et al., 2010a; Ehrenmann and Lefranc, 2011b) where they have been fundamental in the setting up of the standardized definition of contact analysis (Kaas and Lefranc, 2005; Kaas et al., 2008; Ehrenmann et al., 2010a) and of paratope and epitope in crystal structures (Lefranc, 2009; Ehrenmann et al., 2010b).

The IMGT Colliers de Perles are particularly useful in molecular engineering and antibody humanization design based on CDR grafting. Indeed they allow to precisely define the CDR-IMGT and to easily compare the amino acid sequences of FR-IMGT and CDR-IMGT between the mouse (or other species) and the closest human V-DOMAIN (Lefranc, 2009; Ehrenmann et al., 2010b). Analyses performed on humanized therapeutic antibodies underline the importance of a correct delimitation of the CDR regions to be grafted (Magdelaine-Beuzelin et al., 2007). The IMGT Colliers de Perles also allow a comparison to the IMGT Colliers de Perles statistical profiles for the human expressed IGHV, IGKV, and IGLV repertoires. These statistical profiles are based on the definition of 11 IMGT amino acid physicochemical characteristics classes which take into account the hydropathy, volume, and chemical characteristics of the 20 common

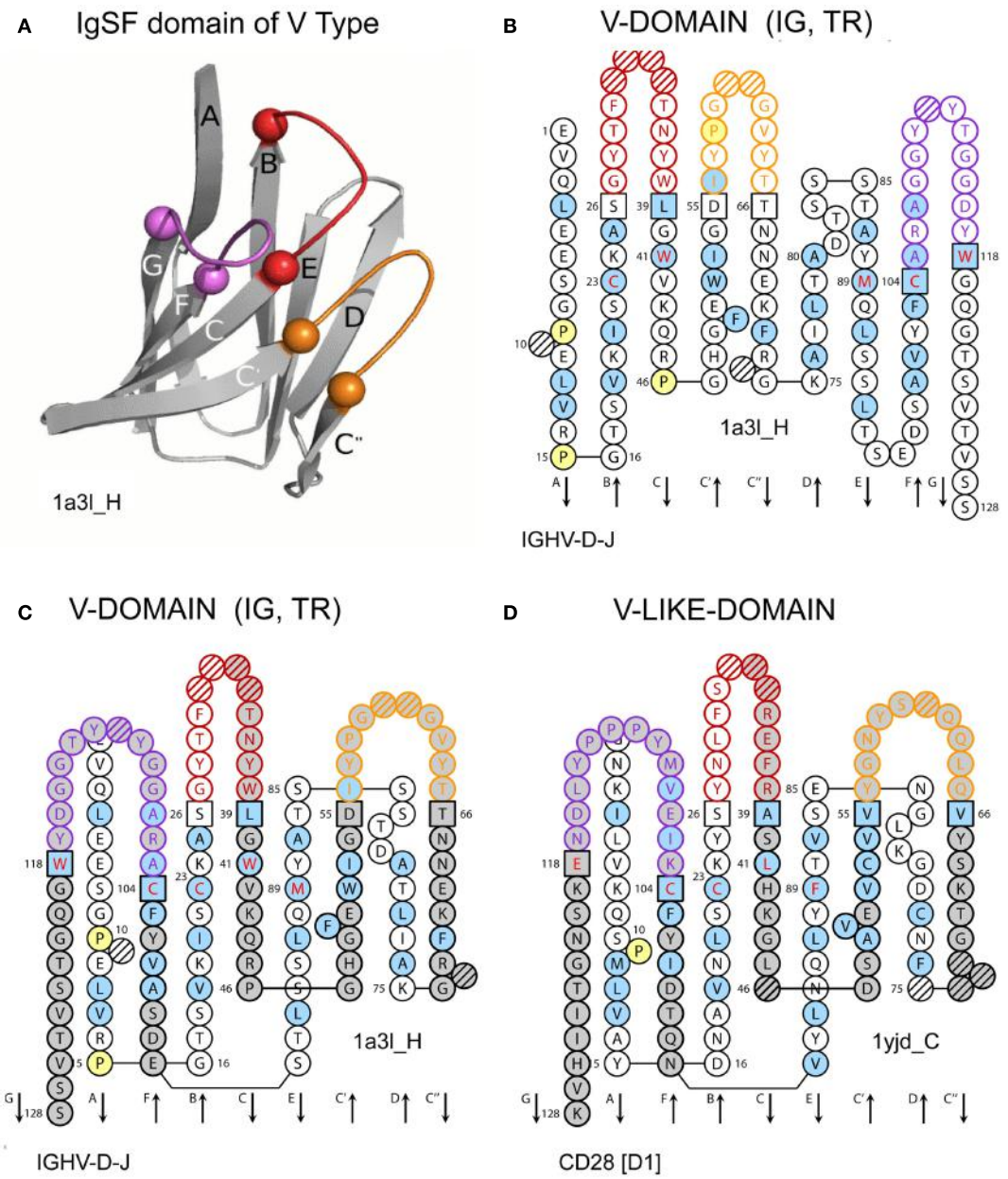


FIGURE 7 | IMGT Collier de Perles for V domain. (A) Ribbon representation of a V-DOMAIN as an example. A similar topology and 3D structure characterize a V-LIKE-DOMAIN. **(B)** and **(C)** V-DOMAIN on one layer and on two layers, respectively (*Mus musculus* VH [8.8.12]). **(D)** V-LIKE-DOMAIN on two layers (*Homo sapiens* CD28 [9.9.13]). Amino acids are shown in the one-letter abbreviation. Positions at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in more than 50% of analyzed sequences are shown online in blue. All proline (P) are shown online in yellow. The loops BC, C′C″ and FG (corresponding to the CDR-IMGT) are limited by amino acids shown in squares (anchor positions), which belong to the neighboring strands (FR-IMGT). BC loops are represented online in red, C′C″ loops in orange and FG loops in purple. Hatched circles or squares correspond to missing positions according to the IMGT unique numbering for V domain (Lefranc et al., 2003). Arrows indicate the direction of the beta strands and their designations in 3D structures. The IMGT Colliers de Perles on two layers show, in the forefront, the GFCC′C″ strands and, in the back, the ABED strands. The chain identifiers to which the domains belong are 1a3l_H for **(A)**, **(B)**, **(C)**, and 1yjd_C for **(D)** [IMGT/3Dstructure-DB (Ehrenmann et al., 2010a; Ehrenmann and Lefranc, 2011b)].

amino acids (Pommié et al., 2004). This comparison is useful to identify potential immunogenic residues at given positions in chimeric or humanized antibodies or to evaluate immunogenicity of therapeutic antibodies.

DISCUSSION

The standardization, the consistency and the reliability of the immunogenetics data in IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) rely on

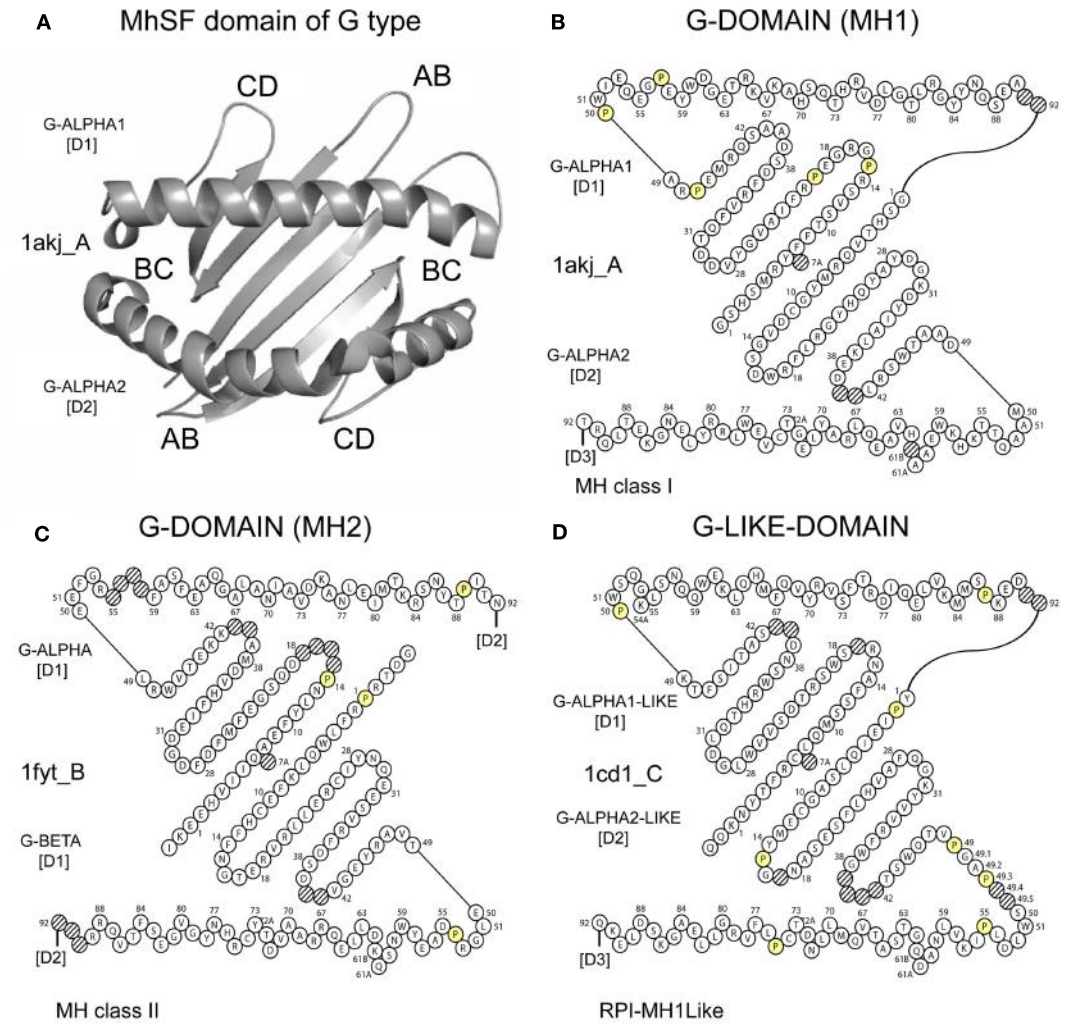


FIGURE 8 | IMGT Collier de Perles for G domain. (A) Ribbon representation of two G-DOMAIN as an example. A similar topology and 3D structure characterize the G-LIKE-DOMAIN. **(B)** G-DOMAIN of MH1: G-ALPHA1 and G-ALPHA2 (*Homo sapiens* HLA-A∗0201). **(C)** G-DOMAIN of MH2: G-ALPHA and G-BETA (*Homo sapiens* HLA-DRA∗0101 and HLA-DRB1∗0101). **(D)** G-LIKE-DOMAIN of RPI-MH1Like: G-ALPHA1-LIKE and G-ALPHA2-LIKE (*Mus musculus* CD1D1). Amino acids are shown in the one-letter abbreviation. Hatched circles correspond to missing positions according to the IMGT unique numbering for G domain (Lefranc et al., 2005c). Note that the N-terminal end of a peptide in the cleft would be on the left hand side. The chain identifiers to which the domains belong are 1akj_A for **(A)** and **(B)**, 1fyt_B for **(C)** and 1cd1_C for **(D)** [IMGT/3Dstructure-DB (Ehrenmann et al., 2010a; Ehrenmann and Lefranc, 2011b)].

IMGT-ONTOLOGY, elaborated since 1989 in order to manage, to share and to represent the immunogenetics knowledge (Giudicelli and Lefranc, 1999; Lefranc et al., 2004, 2005a; 2008; Duroux et al., 2008; Lefranc, 2011a,b,c,d,e,f, 2013).

IMGT-ONTOLOGY has been developed to be used by any scientific domain which deals with immunogenetics. This includes fundamental, medical, veterinary, clinical, pharmaceutical and biotechnological research. Closely related terms have been integrated in some other biological ontologies (**Table 3**). Chain types have been included in NCI Thesaurus, Logical Observation Identifier Names and Codes (LOINC), Molecule role (INOH Protein name/family name ontology) (IMR), National Drug File Reference Terminology (NDRFT). IMGT® standardized labels that describe specifically IG and TR sequences and 3D structures and 64 of the IMGT® standardized labels, in particular those for genomic sequences, have been included in Sequence Ontology (SO; Eilbeck et al., 2005) and in SNP-Ontology. IG and TR gene names were entered in HUGO and NCI Metathesaurus (**Table 3**). These ontologies are available on the NCBO BioPortal (Noy et al., 2009), opening opportunities of mapping with them.

IMGT® standards derived from IMGT-ONTOLOGY concepts allow interoperability between external databases and tools. Interoperability between IMGT®, HGNC, NCBI, Ensembl, and Vega for the concepts of classification has been described (see Interoperability between IMGT, HGNC, and NCBI). The IMGT numbering is integrated in external Web resources: it is proposed, for

Table 3 | Formal IMGT-ONTOLOGY axioms, IMGT-ONTOLOGY concepts, IMGT® standards, and external resources.

Formal IMGT-ONTOLOGY axioms	IDENTIFICATION	DESCRIPTION	CLASSIFICATION	NUMEROTATION
IMGT-ONTOLOGY concepts[a]	Concepts of identification[b] (/1491)	Concepts of description[c]	Concepts of classification[d]	Concepts of numerotation[e,f]
IMGT® standards	IMGT® standardized keywords	IMGT® standardized labels	IMGT® standardized IG and TR gene names	IMGT unique numbering IMGT Colliers de Perles
External resources (ontologies, databases, and tools)	NCI Thesaurus (/1032)	Sequence types and features (SO) (/1109)	HUGO (/1528)	IgBlast
	Logical Observation Identifier Names and Codes (LOINC) (/1350) Molecule role (INOH Protein name/ family name ontology) (IMR) (/1029) National Drug File Reference Terminology (NDRFT) (/1352)	SNP-Ontology (/1058)	NCI Metathesaurus (/1499) HGNC (Bruford et al., 2008) NCBI gene (Maglott et al., 2011) Ensembl (Hubbard et al., 2002) Vega (Wilming et al., 2008)	

(/number) indicates, for ontologies at NCBO BioPortal, the identifiant to be added to http://bioportal.bioontology.org/ontologies.

[a] Giudicelli and Lefranc (1999), Lefranc et al. (2004, 2005a, 2008), Duroux et al. (2008), Lefranc (2013).

[b–f] Lefranc (2011b,c,d,e,f).

example, as domain system numbering in the sequence analysis tool IgBlast[5].

The IMGT® standards generated from IMGT-ONTOLOGY are extensively reused by scientists in very diverse domains for the interpretation of immunogenetics data. The first example is the acknowledgment of the IMGT® gene names as the official nomenclature for IG and TR genes (Wain et al., 2002; Lefranc, 2007, 2008), referenced and recorded in genome sites (NCBI Gene; Maglott et al., 2011). The second example concerns the medical and clinical research which requires a high level of standardization for the results of data analysis in order to take therapeutical decisions: the European Research Initiative on chronic lymphocytic leukemia (CLL) (ERIC) includes 130 laboratories in 26 countries. ERIC has recommended the use of IMGT/V-QUEST (Brochet et al., 2008; Giudicelli et al., 2011), the IMGT® tool for the analysis of IG and TR rearranged sequences, as a reference for determining the rate of IGHV gene mutations, an important prognostic factor for CLL patients (Ghia et al., 2007; Giudicelli and Lefranc, 2008; Langerak et al., 2011). Results provided with the IMGT® standards are integrated in clinical reports (Rosenquist, 2008). The third example is the definition of monoclonal antibodies (mAb, suffix -mab) and fusion proteins for immune applications (FPIA, suffix -cept) of the World Health Organization/International Nonproprietary Name (WHO/INN) programme that are based on the IMGT-ONTOLOGY concepts (Lefranc, 2011g). INN mAb and FPIA have been entered in IMGT/mAb-DB and IMGT/2Dstructure-DB, allowing queries of sequences, 2D structures (or IMGT Collier de Perles) and, if available, 3D structures. The fourth example of great interest for pharmaceutical companies involved in antibody engineering and humanization for therapeutical use is the characterization of the three hypervariable loops (or CDR-IMGT) of an IG or TR variable domain using the IMGT/DomainGapAlign and IMGT/Collier-de-Perles tools. The objective of antibody humanization is to graft the CDR-IMGT of an antibody, usually murine, and of a given specificity onto a human domain framework, thus preserving the original murine antibody specificity while decreasing its immunogenicity (Lefranc, 2009; Ehrenmann et al., 2010b).

IMGT-ONTOLOGY and IMGT® standards ensure the coherency of the IMGT® information system whose data permanently evolve with the most recent advances in science and methodologies. They form a unique and necessary whole for the modeling, the representation and the sharing of the immunogenetics knowledge by both humans and automated resources.

ACKNOWLEDGMENTS

We are grateful to Gérard Lefranc for helpful comments. We thank the IMGT® team, and all the previous collaborators and biocurators for the expertise and constant motivation. IMGT® is an Institutional Academic Member of the International Medical Informatics Association (IMIA). IMGT® is a registered mark of the Centre National de la Recherche Scientifique (CNRS). IMGT® is certified ISO 9001:2008 and has received the National (CNRS, INSERM, CEA, INRA) Bioinformatics Platform labels: RIO in 2001 and IBiSA in 2007. IMGT® is Bioinformatics Platform of ELIXIR, ReNaBi, GDR ACCITH, Cancéropôle GSO, GPTR Sud de France and SFR Biocampus. IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), 5th PCRDT Quality of Life and Management of Living Resources programmes (QLG2-2000-01287) and 6th

[5] http://www.ncbi.nlm.nih.gov/igblast/

PCRDT Information Society Technology (ImmunoGrid, IST-2004-028069) programmes of the European Union (EU). IMGT® was granted access by GENCI to the CINES HPC resources (2010-036029). IMGT® is currently supported by the Ministère de l'Enseignement Supérieur et de la Recherche (MESR), CNRS, Université Montpellier 2, Région Languedoc-Roussillon, Agence Nationale de la Recherche ANR (BIOSYS-06-135457, FLAVORES), and the Labex MabImprove (2011–2020).

REFERENCES

Alamyar, E., Giudicelli, V., Shuo, L., Duroux, P., and Lefranc, M.-P. (2012). IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res.* 8, 1–2.

Ashurst, J. L., Chen, C.-K., Gilbert, J. G. R., Jekosch, K., Keenan, S., Meidl, P., Searle, S. M., Stalker, J., Storey, R., Trevanion, S., Wilming, L., and Hubbard, T. (2005). The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* 33, D459–465.

Bairoch, A., Bougueleret, L., and UniProt Consortium. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* 37, D169–D174.

Bleakley, K., Giudicelli, V., Wu, Y., Lefranc, M.-P., and Biau, G. (2006). IMGT standardization for statistical analyses of T cell receptor junctions: the TRAV-TRAJ example. *In silico Biol. (Gedrukt)* 6, 573–588.

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D. G., Knight, J., Mani, P., Martin, R., Moxon, S. A. T., Paddock, H., Pich, C., Ramachandran, S., Ruef, B. J., Ruzicka, L., Bauer Schaper, H., Schaper, K., Shao, X., Singer, A., Sprague, J., Sprunger, B., Van Slyke, C., and Westerfield, M. (2011). ZFIN: enhancements and updates to the zebrafish model organism database. *Nucleic Acids Res.* 39, D822–D829.

Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36, W503–W508.

Bruford, E. A., Lush, M. J., Wright, M. W., Sneddon, T. P., Povey, S., and Birney, E. (2008). The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.* 36, D445–D448.

Duroux, P., Kaas, Q., Brochet, X., Lane, J., Ginestoux, C., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. *Biochimie* 90, 570–583.

Ehrenmann, F., Giudicelli, V., Duroux, P., and Lefranc, M.-P. (2011). IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb. Protoc.* 2011, 726–736.

Ehrenmann, F., Kaas, Q., and Lefranc, M.-P. (2010a). IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.* 38, D301–D307.

Ehrenmann, F., Duroux, P., Giudicelli, V., and Lefranc, M.-P. (2010b). "Standardized sequence and structure analysis of antibody using IMGT®," in *Antibody Engineering,* Vol. 2, eds R. Kontermann and S. Dübel (Berlin: Springer-Verlag), 11–31.

Ehrenmann, F., and Lefranc, M.-P. (2011a). IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb. Protoc.* 2011, 737–749.

Ehrenmann, F., and Lefranc, M.-P. (2011b). IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb. Protoc.* 2011, 750–761.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.* 6, R44.

Garapati, V. P., and Lefranc, M.-P. (2007). IMGT Colliers de Perles and IgSF domain standardization for T cell costimulatory activatory (CD28, ICOS) and inhibitory (CTLA4, PDCD1 and BTLA) receptors. *Dev. Comp. Immunol.* 31, 1050–1072.

Ghia, P., Stamatopoulos, K., Belessi, C., Moreno, C., Stilgenbauer, S., Stevenson, F., Davi, F., and Rosenquist, R. (2007). ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leukemia* 21, 1–3.

Giudicelli, V., Brochet, X., and Lefranc, M.-P. (2011). IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb. Protoc.* 2011, 695–715.

Giudicelli, V., Chaume, D., Jabado-Michaloud, J., and Lefranc, M.-P. (2005a). Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. *Stud. Health Technol. Inform.* 116, 3–8.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2005b). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33, D256–D261.

Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., and Lefranc, M.-P. (2006). IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.* 34, D781–D784.

Giudicelli, V., and Lefranc, M.-P. (1999). Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics* 15, 1047–1054.

Giudicelli, V., and Lefranc, M.-P. (2008). "IMGT® standardized analysis of immunoglobulin rearranged sequences," in *Immunoglobulin Gene Analysis in Chronic Lymphocytic Leukemia*, eds P. Ghia, R. Rosenquist, and F. Davi (Milan: Wolters Kluwer Health), 33–52.

Giudicelli, V., and Lefranc, M.-P. (2011). IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb. Protoc.* 2011, 716–725.

Greenaway, H. Y., Kurniawan, M., Price, D. A., Douek, D. C., Davenport, M. P., and Venturi, V. (2009). Extraction and characterization of the rhesus macaque T-cell receptor beta-chain genes. *Immunol. Cell Biol.* 87, 546–553.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquis.* 5, 199–220.

Guarino, N. (1997). Understanding, building and using ontologies. *Int. J. Hum. Comput. Stud.* 46, 293–310.

Guarino, N., and Giaretta, P. (1995). "Ontologies and knowledge bases: towards a terminological clarification," in *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing,* ed. N. Mars (Amsterdam, NL: IOS Press), 25–32.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41.

Kaas, Q., Duprat, E., Tourneur, G., and Lefranc, M.-P. (2008). "IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes," in *Immunoinformatics, Immunomics Reviews, Series of Springer Science and Business Media LLC,* eds C. Schoenbach, S. Ranganathan, and V. Brusic (New York: Springer), 19–49.

Kaas, Q., Ehrenmann, F., and Lefranc, M.-P. (2007). IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? *Brief. Funct. Genomic. Proteomic.* 6, 253–264.

Kaas, Q., and Lefranc, M.-P. (2005). T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In silico Biol. (Gedrukt)* 5, 505–528.

Kaas, Q., and Lefranc, M.-P. (2007). IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Curr. Bioinform.* 2, 21–30.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer,

S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Lane, J., Duroux, P., and Lefranc, M.-P. (2010). From IMGT-ONTOLOGY to IMGT/LIGMotif: the IMGT standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC Bioinformatics* 11, 223. doi:10.1186/1471-2105-11-223

Langerak, A. W., Davi, F., Ghia, P., Hadzidimitriou, A., Murray, F., Potter, K. N., Rosenquist, R., Stamatopoulos, K., and Belessi, C. (2011). Immunoglobulin sequence analysis and prognostication in CLL: guidelines from the ERIC review board for reliable interpretation of problematic cases. *Leukemia* 25, 979–984.

Lefranc, M. P. (1997). Unique database numbering system for immunogenetic analysis. *Immunol. Today* 18, 509.

Lefranc, M. P. (1999). The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *Immunologist* 7, 132–136.

Lefranc, M.-P. (2000a). "Nomenclature of the human immunoglobulin genes," in *Current Protocols in Immunology*, eds J. E. Coligan, B. E. Bierer, D. E. Margulies, E. M. Shevach, and W. Strober (Hoboken, NJ: John Wiley and Sons, Inc), A.1P.1–A.1P.37.

Lefranc, M.-P. (2000b). "Nomenclature of the human T cell receptor genes," in *Current Protocols in Immunology*, eds J. E. Coligan, B. E. Bierer, D. E. Margulies, E. M. Shevach, and W. Strober (Hoboken, NJ: John Wiley and Sons, Inc), A.1O.1–A.1O.23.

Lefranc, M.-P. (2005). IMGT, the international ImMunoGeneTics information system: a standardized approach for immunogenetics and immunoinformatics. *Immunome Res.* 1, 3.

Lefranc, M.-P. (2007). WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. *Immunogenetics* 59, 899–902.

Lefranc, M.-P. (2008). WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev. Comp. Immunol.* 32, 461–463.

Lefranc, M.-P. (2009). "Antibody databases and tools: The IMGT® experience," in *Therapeutic Monoclonal Antibodies: From Bench to Clinic*, ed. Z. An (Hoboken, NJ: John Wiley and Sons, Inc), 91–114.

Lefranc, M.-P. (2011a). IMGT, the international ImMunoGeneTics information system. *Cold Spring Harb. Protoc.* 2011, 595–603.

Lefranc, M.-P. (2011b). From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb. Protoc.* 2011, 604–613.

Lefranc, M.-P. (2011c). From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb. Protoc.* 2011, 614–626.

Lefranc, M.-P. (2011d). From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb. Protoc.* 2011, 627–632.

Lefranc, M.-P. (2011e). IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of, IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb. Protoc.* 2011, 633–642.

Lefranc, M.-P. (2011f). IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb. Protoc.* 2011, 643–651.

Lefranc, M.-P. (2011g). Antibody nomenclature: from IMGT-ONTOLOGY to INN definition. *MAbs* 3, 1–2.

Lefranc, M.-P. (2013). "IMGT-ONTOLOGY," in *Encyclopedia of Systems Biology*, eds W. Dubitzky, O.Wolkenhauer, K.-H. Cho, and H. Yokota (New York: Springer) (in press).

Lefranc, M.-P., Clement, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D., and Lefranc, G. (2005a). IMGT-Choreography for immunogenetics and immunoinformatics. *In silico Biol. (Gedrukt)* 5, 45–60.

Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piédade, I., Rouard, M., Foulquier, E., Thouvenin, V., and Lefranc, G. (2005b). IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.* 29, 185–203.

Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A., and Lefranc, G. (2005c). IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev. Comp. Immunol.* 29, 917–938.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bosc, N., Folch, G., Guiraudou, D., Jabado-Michaloud, J., Magris, S., Scaviner, D., Thouvenin, V., Combres, K., Girod, D., Jeanjean, S., Protat, C., Yousfi-Monod, M., Duprat, E., Kaas, Q., Pommié, C., Chaume, D., and Lefranc, G. (2004). IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In silico Biol. (Gedrukt)* 4, 17–29.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., and Duroux, P. (2009). IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 37, D1006–D1012.

Lefranc, M.-P., Giudicelli, V., Regnier, L., and Duroux, P. (2008). IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief. Bioinform.* 9, 263–275.

Lefranc, M.-P., and Lefranc, G. (2001a). *The Immunoglobulin FactsBook*. London: Academic Press.

Lefranc, M.-P., and Lefranc, G. (2001b). *The T Cell Receptor FactsBook*. London: Academic Press.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27, 55–77.

Letovsky, S. I., Cottingham, R. W., Porter, C. J., and Li, P. W. (1998). GDB: the Human Genome Database. *Nucleic Acids Res.* 26, 94–99.

Magdelaine-Beuzelin, C., Kaas, Q., Wehbi, V., Ohresser, M., Jefferis, R., Lefranc, M.-P., and Watier, H. (2007). Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. *Crit. Rev. Oncol. Hematol.* 64, 210–225.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 35, D26–D31.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez Gene:

gene-centered information at NCBI. *Nucleic Acids Res.* 39, D52–D57.

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. (2000). NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* 28, 126–128.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M.-A., Smith, B., and the NCBO team. (2012). The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* 19, 190–195.

Noy, N. F., Crubezy, M., Fergerson, R. W., Knublauch, H., Tu, S. W., Vendetti, J., and Musen, M. A. (2003). Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu. Symp. Proc.* 953.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 37, W170–W173.

Pappalardo, F., Lefranc, M.-P., Lollini, P.-L., and Motta, S. (2010). A novel paradigm for cell and molecule interaction ontology: from the CMM model to IMGT-ONTOLOGY. *Immunome Res.* 6, 1.

Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G., and Lefranc, M.-P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.* 17, 17–32.

Prabakaran, P., Chen, W., Singarayan, M. G., Stewart, C. C., Streaker, E., Feng, Y., and Dimitrov, D. S. (2012). Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64, 337–350.

Rosenquist, R. (2008). "How to report IG sequence data in clinical routine: cases difficult to categorize," in *Immunoglobulin Gene Analysis in Chronic Lymphocytic Leukemia*, eds P. Ghia, R. Rosenquist, and F. Davi (Milan: Wolters Kluwer Health), 113–124.

Ruiz, M., and Lefranc, M.-P. (2002). IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* 53, 857–883.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science* 291, 1304–1351.

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., and Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics* 79, 464–470.

Wain, H. M., Lush, M. J., Ducluzeau, F., Khodiyar, V. K., and Povey, S. (2004). Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.* 32, D255–D257.

Wilming, L. G., Gilbert, J. G. R., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J. L. (2008). The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* 36, D753–D760.

Yousfi Monod, M., Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004). IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* 20(Suppl. 1), i379–i385.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 February 2012; accepted: 24 April 2012; published online: 23 May 2012.

Citation: Giudicelli V and Lefranc M-P (2012) IMGT-ONTOLOGY 2012. Front. Gene. ***3****:79. doi: 10.3389/fgene.2012.00079*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 28 May 2012
doi: 10.3389/fgene.2012.00087

Three ontologies to define phenotype measurement data

Mary Shimoyama[1]*, Rajni Nigam[1], Leslie Sanders McIntosh[2], Rakesh Nagarajan[2], Treva Rice[2], D. C. Rao[2] and Melinda R. Dwinell[1]

[1] *Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, USA*
[2] *Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA*

Edited by:
John Hancock, Medical Research Council, UK

Reviewed by:
Philippe Rocca-Serra, Oxford e-Research Centre, UK
Peter N. Robinson, Charité, Germany

***Correspondence:**
Mary Shimoyama, Department of Surgery, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA.
e-mail: shimoyama@mcw.edu

Background: There is an increasing need to integrate phenotype measurement data across studies for both human studies and those involving model organisms. Current practices allow researchers to access only those data involved in a single experiment or multiple experiments utilizing the same protocol. **Results:** Three ontologies were created: Clinical Measurement Ontology, Measurement Method Ontology and Experimental Condition Ontology. These ontologies provided the framework for integration of rat phenotype data from multiple studies into a single resource as well as facilitated data integration from multiple human epidemiological studies into a centralized repository. **Conclusion:** An ontology based framework for phenotype measurement data affords the ability to successfully integrate vital phenotype data into critical resources, regardless of underlying technological structures allowing the user to easily query and retrieve data from multiple studies.

Keywords: ontology, phenotype

BACKGROUND

The quest to link characteristics of an individual or organism to genetic structures dates to the mid-1800s and the work of Gregor Mendel (Sorsby, 1965). In the past 20 years, a great deal of progress has been made in identifying, naming, and standardizing the information about genetic structures. The International Nucleotide Sequence Database Collaboration[1] was created to develop standard formats for genomic data to integrate data generated at multiple laboratories using a variety of technologies resulting in public databases housed at the National Center for Biotechnology Information (NCBI; Sayers et al., 2010), the DNA Databank of Japan (Kaminuma et al., 2010), and the European Bioinformatics Institute (EBI; Goujon et al., 2010). This integration of data has led to the development of numerous data mining, presentation, and analysis tools and provides a platform for comparisons of genetic and genomic structures across species. Unfortunately, a similar development in data standards and integration has not occurred for the characteristics of an individual or organism scientists wish to link to these structures. The potential value of integrating phenotype data from multiple sources (e.g., different laboratories or studies, varying techniques to measure similar phenotypes, multiple populations, or strains of a particular organism) is enormous. The power to identify novel genes associated with human disease and the role a gene plays in disease is greatly increased with clearly defined phenotype information and the inclusion of the environmental and experimental context (Butte and Kohane, 2006). However, most phenotype data is gathered or generated without thought to integrating the results with those from other studies even within the same laboratory or program, creating barriers to integrating and comparing results reported in publications. In both animal and human physiological and disease studies, there has been a long tradition of designing new protocols and adopting evolving best practices available at the time the study is launched. As a result, the same basic information gets collected differently across protocols. This leads to a common belief that each study is unique and cannot be compared to any other for anything more than the most general elements. Moreover, the data sets and study information are structured in such a way that, often, only those who are intimately familiar with the study understand the full depth of the data; this includes details such as the measurement methods used and the experimental conditions imposed. For most researchers, ferreting out this information from different studies requires extensive time and effort, as is generally experienced by *post hoc* collaborations among multiple studies. Even when these details are published, they are often described in widely different ways without full inclusion of details making comparisons across studies not only difficult but sometimes impossible.

Variations in experimental conditions, population, age, and study design all contribute to the difficulty in comparing phenotype data from multiple sources. For example, the comparison of blood pressure measured in different laboratories or programs can be impacted by the way in which blood pressure is measured (e.g., direct measurement via catheter in artery, telemetry, blood pressure cuff), the experimental conditions imposed as part of the study (e.g., low salt/high salt diet, exercise, oxygen levels), surgical manipulations (e.g., removal of a kidney), gender, and age. One approach to aggregating and integrating phenotype data would be to develop standard phenotyping protocols to be followed by all researchers. However, standardizing the methods used for phenotyping protocols has significant drawbacks. Many would see it as impractical since each researcher is testing fairly unique hypotheses which cannot be easily investigated by using a set protocol. Additionally, not all laboratories measure phenotypes using the same assays, nor do all investigators agree on one perfect method to measure each phenotype. Any movement

[1] http://www.insdc.org/

toward this type of standardization would take years before results were evident, keeping existing data resources inaccessible. A more practical approach is to develop a method using ontologies and standardized data formats to integrate phenotype measurement data sets.

A number of groups have focused on standardizing biological information through standardized vocabularies and ontologies. Ontologies are hierarchically structured vocabularies of terms and relationships that are clearly defined and designed to represent and communicate information about a particular scientific domain (**Figure 2**). The entities and concepts represented by the terms in the lower nodes are assumed to inherit the properties and qualities of those of nodes higher up the branch. The National Institutes of Health, in recognition of the utility of ontologies and the need for more ontologies to represent biological concepts, provided funding for the creation of the National Center for Biomedical Ontology (NCBO)[2] in 2006 (Rubin et al., 2006). There are currently 242 ontologies cataloged at NCBO including several which focus on phenotypes.

MAMMALIAN PHENOTYPE ONTOLOGY

The Mammalian Phenotype (MP) Ontology was initially created for annotating gene alleles at the Mouse Genome Informatics (MGI) database (Smith et al., 2005). For the MP ontology, "phenotype refers to the observable morphological, physiological, and behavioral characteristics of an individual in the context of the environment" (Smith and Eppig, 2009). Because MP was designed to be used with mouse knockouts, mutations, and other types of alleles, there is an underlying assumption of a comparison to the trait exhibited by a mouse with the genetic background from which the allele has been constructed or a comparison to a normal or "wild type" trait. Thus the terms often contain words such as "abnormal," "increased," or "decreased" with the implication of "relative to" an assumed observation. The actual measured values for observed traits are not connected to these annotations. MP follows the open-source Open Biological and Biomedical Ontology (OBO) file format and is organized on the Directed Acyclic Graph (DAG) structure with the highest nodes related to physiological systems such as cardiovascular, immune system as well as behavioral, life span, and cellular phenotypes. Each physiological system node is followed by a basic division into physiological and morphological phenotype branches. MGI currently has over 41,000 genotypes annotated with MP terms for a total of more than 193,000 annotations. MP is considered a pre-coordinated term ontology since both the entity (i.e., anatomical site or physiological process) and the quality of it (i.e., abnormal, increased, decreased) are included in the term. MP is also being used for the EuroPhenome project to annotate mutant mouse phenotype data generated using standard phenotyping platforms (Morgan et al., 2010). The advantages in terms of annotation are significant since curators only have to search a single ontology and has terms that more closely mimic those seen in literature and commonly used in laboratory settings.

[2]http://bioontology.org/

PHENOTYPE AND TRAIT ONTOLOGY

Another approach to phenotype ontologies has been the Phenotype and Trait Ontology (PATO) project[3] (Gkoutos et al., 2004). Unlike MP which is considered a pre-coordinated term phenotype ontology, PATO uses the EQ approach (entity + quality; Gkoutos et al., 2004; Smith and Eppig, 2009). Thus PATO presents terms related to qualities and attributes that are then linked to terms from other ontologies such as anatomy ontologies to describe phenotypic characteristics. Thus, "big ears" would be represented by the term "increased size" from PATO and the word "ear" from an anatomy ontology (Mungall et al., 2010). One of the advantages of this approach is the re-use of existing ontologies such as the anatomy ontology. Representing morphological traits through the use of anatomy ontologies and the qualities described in PATO is relatively straight forward. This is one reason that resources housing data for some organisms such as drosophila and zebrafish have found this approach useful (Mungall et al., 2010); the majority of their reported traits and phenotypes are morphological in nature. However, the representation of physiological traits and specific clinical measurement types is more problematic (Smith and Eppig, 2009). First, there is not necessarily a single ontology that adequately represents the physiological trait corresponding with the quality expressed by PATO; and second, a single EQ term may not adequately express the phenotype observed. For example while the morphological trait of "big ears" is relatively easy to represent by EQ, an MP term of "abnormal cochlear outer hair cell electromotility" provides a greater challenge. The disadvantages of the PATO approach for annotation are also significant. Curators would have to browse multiple ontologies to create a term on the fly and this approach creates terms and phrases that sometimes are stilted or not commonly used in the literature or laboratory settings.

HUMAN PHENOTYPE ONTOLOGY

The Human Phenotype Ontology (HPO) was developed in part to address the shortcomings of information presented in the Online Mendelian Inheritance in Man (OMIM) database[4] (Amberger et al., 2009). OMIM has traditionally been the most commonly used resource for information on genetic diseases. Unfortunately, the information housed there is in free text format, making it difficult to mine computationally because of the non-standard way in which traits and abnormalities are described. For instance, OMIM uses the synonymous descriptions "generalized amyotrophy," "generalized muscular atrophy," and "muscular atrophy, generalized" so even simple searches may not return the results a user desires. While a human reader going through the free text of multiple entries will recognize similar meanings, computers will not. The initial version of HPO was created using the information at OMIM in an effort to merge synonyms and create links and relationships among the terms and concepts. This initial structure has been expanded and refined through manual curation of information from a variety of sources and consistent development of definitions and relationships (Robinson and Mundlos, 2010). As

[3]http://www.bioontology.org/wiki/index.php/PATO:Main_Page
[4]http://www.ncbi.nlm.nih.gov/omim

with MP, the emphasis has been on phenotypes which diverge from the normal or expected and disease states and terms are pre-constructed as with the MP.

Ontologies such as MP, PATO, and HPO were originally designed for use in simple annotations to a single data (e.g., gene product or allele) or an individual and the term was expected to appear alone in the annotation with the minimal accompanying information of an evidence code indicating level of experimental evidence to support the annotation and the reference from which the annotation was made. The existing phenotype ontologies were not developed to be attached to actual measurement values but to indicate a state or characteristic observed relative to that which has been determined to be "normal" or "wild type," or relative to that exhibited by an individual with a known genotype. Information on experimental conditions and measurement assays used are vital parts of the phenotype record and the use of multiple ontologies to represent these has been advocated as a way to accomplish this (Shimoyama et al., 2005; Hancock et al., 2007). Clearly, developing separate ontologies for the elements of phenotype measurement, method of measurement, and conditions under which the measurement was made along with provisions for additional information on actual values, duration of conditions, and so on, will allow these aspects of the phenotype record to be linked. Database structures which allow re-use of information and multiple associations will facilitate data integration, data mining, and data presentation.

In this paper, we present three ontologies created to standardize phenotype measurement records for use in human studies and those using laboratory animals: Clinical Measurement Ontology (CMO), Measurement Method Ontology (MMO), Experimental Condition Ontology.

MATERIALS AND METHODS

The standard elements of phenotype measurement records were identified as: (1) what was measured, (2) how it was measured, and (3) under what conditions it was measured. Ontologies were developed to standardize each of these elements (**Figure 1**) and include (1) CMO, (2) MMO, and (3) Experimental Conditions Ontology (XCO).

The ontologies are available through the NCBO Bioportal[5], and at the Rat Genome Database in ftp files[6]. The ontologies undergo revisions and updates for both consistencies in format and to extend the breadth and depth of coverage as new measurement records are added. The ontologies were developed using the OBO format and the OBO Edit tool (Day-Richter et al., 2007) available at the NCBO, through its Foundry project (Smith et al., 2007). While developed in OBO, developments in OBO to Web Ontology Language (OWL) mapping tools should facilitate conversion to this other highly used ontology format. The major ontology

[5]http://bioportal.bioontology.org/
[6]http://rgd.mcw.edu/pub/ontology

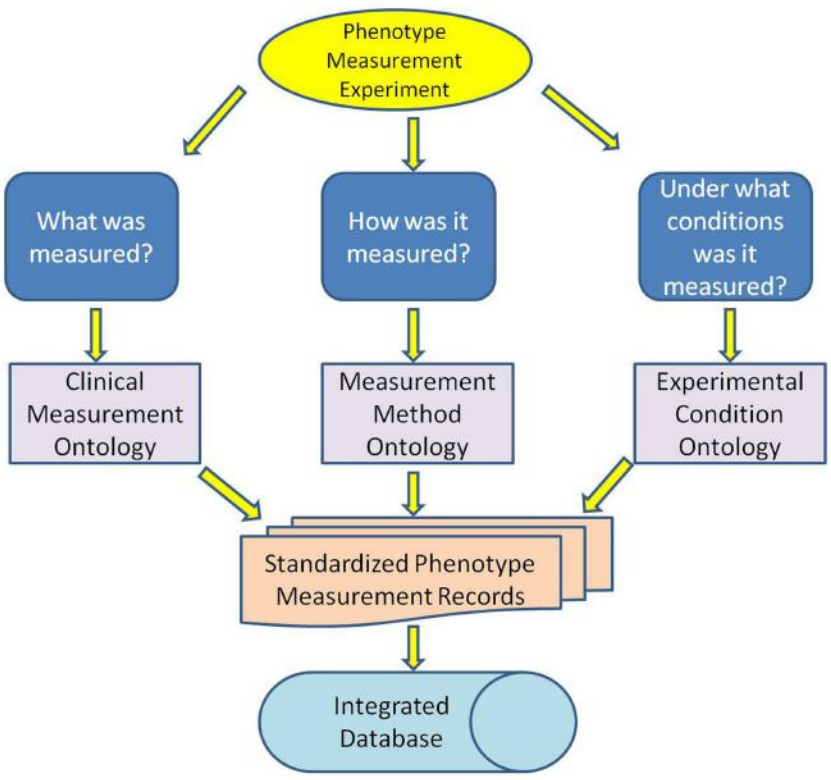


FIGURE 1 | Three ontologies were developed to standardize the three elements of a measurement record: what was measured, how it was measured and under what conditions it was measured.

development tools, OBO Edit for OBO and Protege for OWL now offer widgets that facilitate conversion from one file format to the other (see text footnote 2). The OBO Relation Ontology (RO) was used to create consistency in relationship representations (Smith et al., 2007). These ontologies follow the form of DAGs in which there is a set of nodes with edges forming the linkage between nodes (Robinson and Mundlos, 2010). The nodes are the terms in the ontology with the edges representing the relationships between nodes and the overall visualization of such ontologies resembles branches. In DAGs, the edges or relationships are one way, moving from one node to another, and they do not cycle back. The general relationship pattern in many of these ontologies is a movement from the more general (higher nodes) to the more specific (lower nodes). The entities and concepts represented by the terms in the lower nodes are assumed to inherit the properties and qualities of those of nodes higher up the branch.

The development of these ontologies has included cross references with other ontologies when an exact match of the entity exists. For example, relationships were created with ChEBI in the Experimental Condition Ontology and to the Electrocardiography Ontology (ECG) exist in the CMO. These relationships were created manually and are not used to create cross products. Cross referencing to other ontologies, such as the Cell Ontology, Evidence Code Ontology, and other ontologies used for the reporting of phenotypes, will continue with both manual and semi-automated methods as the ontologies are extended.

CLINICAL MEASUREMENT ONTOLOGY

The CMO provides the standardized vocabulary necessary to indicate the type of measurement made to assess a trait. For the purposes of this project and these ontologies, trait and clinical measurement are defined as follows:

Trait

A physiological or morphological state or property found in all members of a species. Traits can be described or assessed quantitatively (numerically) or qualitatively based on the results of an appropriate form of measurement. The assessment of the trait is not equivalent to the trait itself. Traits exist even when they are not assessed or measured. Often multiple forms of measurement are used to assess a single property or state.

Measurement

The act or result of the act of assessing a morphological or physiological state or property in a single individual or group of individuals and assigning a quantitative or qualitative value. A measurement does not exist until it is performed or taken. Often a single measurement can be used to assess multiple properties or states, sometimes in conjunction with other measurements.

For example, all humans have intelligence or mental capacity, but not all human individuals have an IQ because it has not yet been measured. Similarly, all humans have a body mass but they do not all have a body weight because it has not yet been measured.

Each term in the CMO describes a distinct type of measurement used to assess one or more traits. The terms are arranged in a hierarchical structure of classes so that lower classes are subclasses of higher classes in the branch (**Figure 2**).

This represents an "is_a" type relationship so that a lower term "is_a" subclass of a higher term. Thus, blood cell measurement "is_a" blood measurement and complete blood cell count "is_a" blood cell measurement. The measurements in the ontology are primarily organized on the highest level according to the body system in which the measurement is made. Trait areas were targeted for ontology development based on the availability and extent of data in large scale rat phenotyping projects, published rat literature with phenotype measurements and targeted human epidemiological studies. Ontology development began with the identification of clinical measurements used to assess targeted traits, with terms and definitions being created and relationships among terms being set (**Figure 3**).

Existing ontologies at NCBO were reviewed for associated terms and definitions. Because the clinical measurements in the targeted sources may be limited, additional literature including medical and physiological textbooks, laboratory manuals, and published research literature were reviewed to ensure completeness in the ontology. For example, to assess kidney mass the data source may only use right kidney weight. Further review of a variety of sources reveals that other typical measurements would include left kidney weight, weight of both kidneys, kidney weight expressed as a percentage of body weight which are also often used to assess the trait of kidney mass so these were also added to the ontology. For every clinical measurement term created, associated measurement method terms and experimental condition terms were created based on data in the originating sources as well as the review of additional literature and existing ontologies. There are currently 523 terms in the CMO for measurement types ranging from morphological to physiological for blood, cardiovascular, respiratory, renal, and other systems as well as for growth, reproduction, consumption, tumors, and tissue composition.

MEASUREMENT METHOD ONTOLOGY

A critical element in the description of a phenotype is the measurement method used. Several types of methods are commonly used to measure such things as blood pressure resulting in differing clinical measurement values so the inclusion of method as part of the measurement reading is necessary for the integration of data from multiple studies. The MMO is designed to provide this information. As described above, this ontology was developed in parallel with the CMO as trait areas are targeted. The MMO is organized by the underlying principle or mechanism of the method (**Figure 4**) with two major branches, "*ex vivo* method" and "*in vivo* method." Methods were identified from protocol descriptions and data labels from the targeted data sources. As with the CMO, for completeness in the ontology, additional sources of method information such as vendors' catalogs, laboratory manuals, and published literature were reviewed for associated methods. Thus, if one of the protocols for one of the originating data sources indicated that a balloon tipped catheter was used to measure blood pressure, a quick review of a variety of publications revealed that the basic category is vascular indwelling catheter with a variety of types including fluid filled catheter, intravascular electromagnetic flow sensor, and transducer tipped catheter. There are currently 195 terms in the MMO.

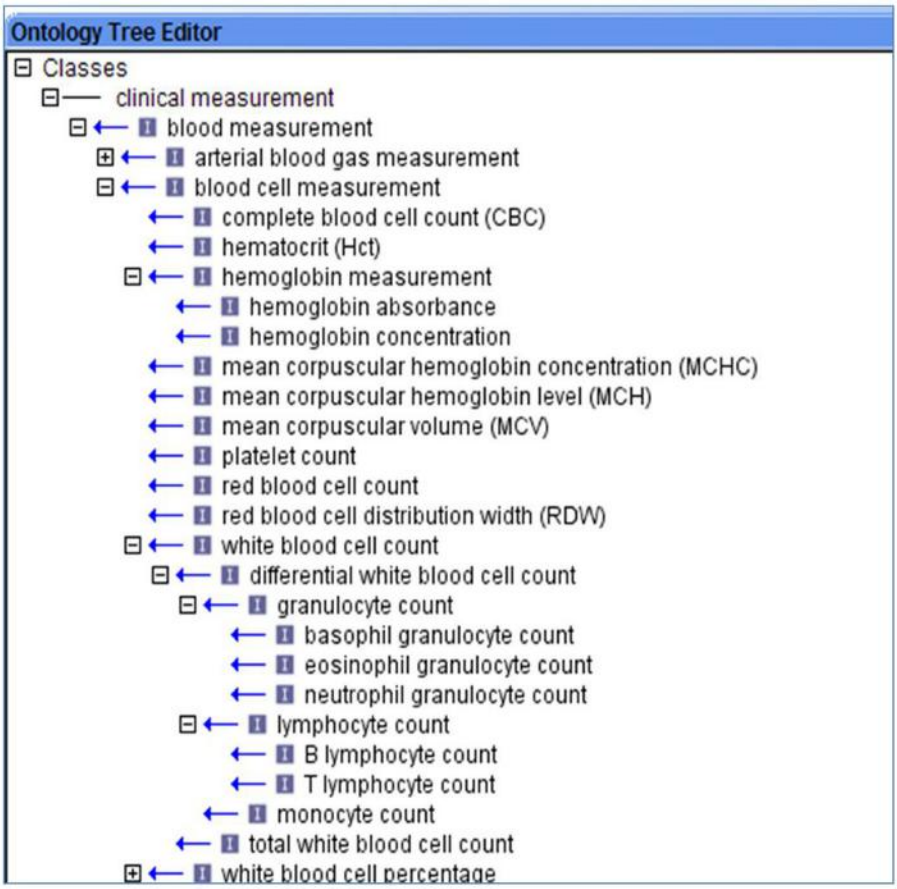


FIGURE 2 | The Clinical Measurement Ontology is presented in a hierarchical structure with classes lower down a branch being subclasses of those above with an "is_a" relationship.

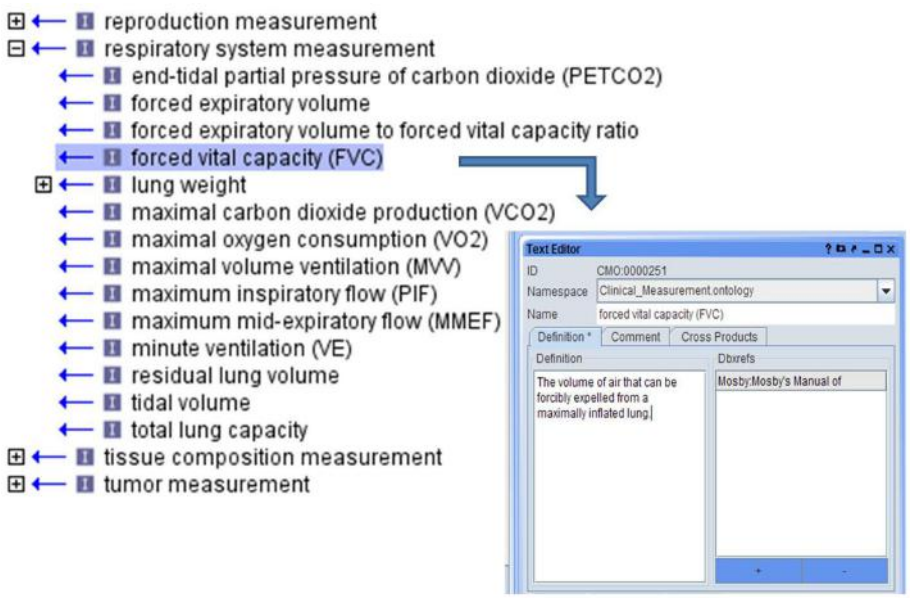


FIGURE 3 | Each CMO term was created as phenotype domains addressed with appropriate definitions for each term.

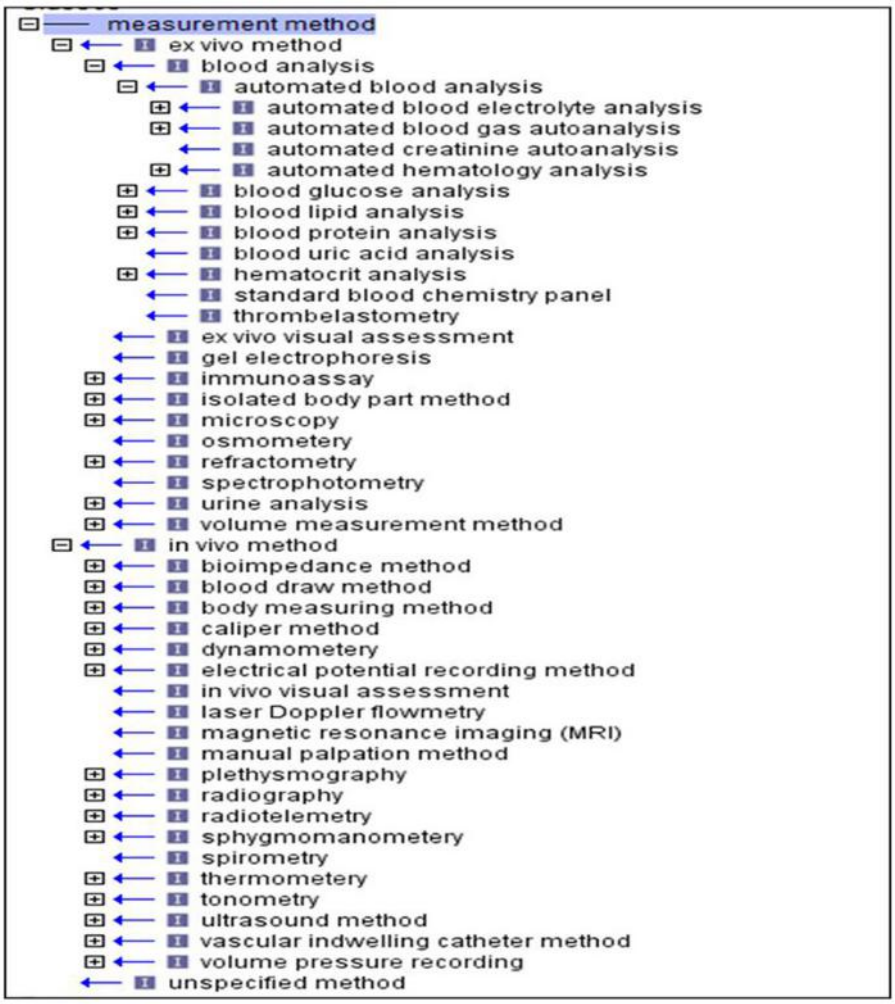


FIGURE 4 | The Measurement Method Ontology structure is based on two major branches, "*ex vivo*" and "*in vivo*" and the underlying mechanism or technique used in the method.

EXPERIMENTAL CONDITION ONTOLOGY

While many phenotype measurements are made under baseline conditions, changes to diet, atmosphere, activity level, and other conditions are common aspects of phenotype experiments. Often this information is added as part of the phenotype label in the individual laboratory's database or only included as part of a lengthy text protocol. Creation of standardized terminology and format for presenting this information with phenotype measurements is crucial to the integration of phenotype data from multiple datasets. An XCO was created to provide standardization and structure for this important information (**Figure 5**).

In addition, to the "is_a" relationship, in certain areas a "part_of" relationship is utilized so that parts of a whole can be described as in "air oxygen content" is "part_of" "atmosphere composition." The ontology was designed so that conditions related to existing ontologies such as those involving chemicals or drugs represented in ChEBI (Degtyarenko et al., 2009), follow the structure and terminology of these ontologies and provide appropriate linkages through identifiers (**Figure 5**). Initial emphasis was on the conditions used in the targeted data sets and expanded to those conditions most commonly used in experiments involving the targeted trait domains. Structural provisions in the database structures of projects using the ontology provide ordinality information to indicate whether multiple conditions were simultaneous or sequential. Use of this ontology in annotating phenotype data allows users to retrieve multiple, disparate phenotype information in which similar experimental conditions were imposed. There are currently 110 terms in the XCO.

DATA INTEGRATION

The three ontologies have been used to integrate multiple data sets for two major projects, one involving human data and the other involving rat data. The Cardiovascular Ontologies and Vocabularies in Epidemiological Research project was designed to integrate demographic and phenotype measurement data from three family blood pressure studies, Hypertension Genetic Epidemiology Network (HyperGEN; Williams et al., 2000); Genetic Epidemiology Network of Salt Sensitivity (GenSalt; Gu et al., 2007);

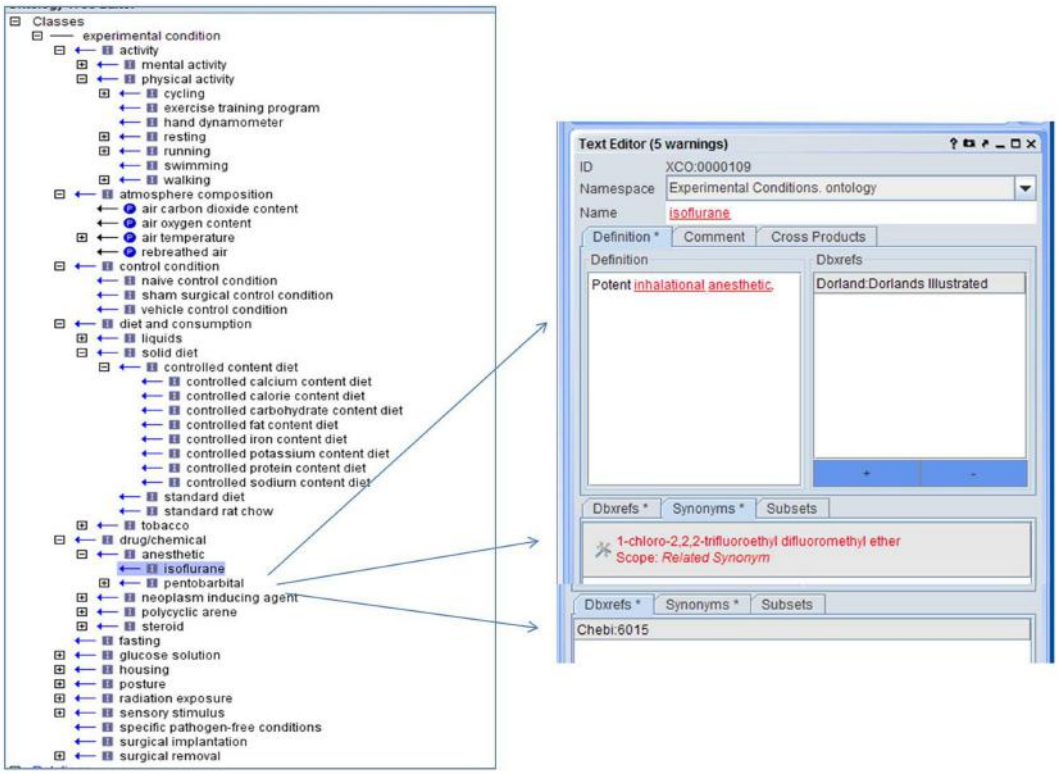


FIGURE 5 | The Experiment Condition Ontology is structured by type of condition with both "is_a" and "part_of" relationships with links to identifiers found in other ontologies.

and HEalth, RIsk factors exercise Training And GEnetics (HERITAGE; Bouchard et al., 1995). While all three studies focused on cardiovascular disease and associated risk factors, they were disparate in the types of interventions used, variety of measurements taken and the methods used to make the measurements. Measurements related to blood pressure, blood chemistry and lipid levels, body weight and body fat were included as well as interventions ranging from sodium controlled diets to exercise. Invasive, non-invasive and imaging techniques were also used. The HyperGEN study was designed to characterize the genes influencing hypertension by recruiting hypertensive sibships (i.e., each participant with two or more hypertensive sibs) from across multiple field centers and ethnic groups. The GenSalt study is an intervention study of the genetic and environmental factors related to dietary sodium and potassium effects on blood pressure in rural Chinese families. The HERITAGE study is an intervention study designed to assess genetic and environmental factors underlying the effects of endurance exercise training on several cardiorespiratory and cardiovascular disease risk factors. The CMO, MMO, and Experimental Condition Ontology were used to map data elements from each of the studies to a common format for integration into a single resource[7]. To date, 16 phenotype classes with records for 8,778 subjects have been integrated for the three studies and made available at the website. Additionally, all variables across the studies are being mapped and modeled with their associated ontology terms to facilitate querying and access to raw data fields of interest; to date 11 classes have been created representing over 100 phenotype measurements.

The rat PhenoMiner project[8] (**Figure 6**) also has used the three ontologies to define data formats and standards for integrating rat phenotype measurement data from a variety of sources including two large scale phenotyping projects and published literature. PhysGen Program for Genomic Applications[9] (Kwitek et al., 2006), one of the large scale phenotyping projects was designed to conduct high throughput phenotype screening for a targeted set of inbred strains, as well as consomic and mutant strains. The screens involved hundreds of different types of phenotype measurements for heart, lung, renal, vascular, and blood function under baseline conditions as well as varying diet, atmosphere, and activity conditions. Data was organized, stored, and presented by protocol so even though some similar measurements such as weight or blood pressure were measured in multiple protocols, the data was not integrated across protocols. The National BioResource Project for the Rat in Japan (Serikawa et al., 2009) was the second large scale rat phenotyping project. Phenotype screens for body weight, activity, behavior, blood pressure, blood chemistry, urine analysis,

[7]http://cover.wustl.edu/Cover/

[8]http://rgd.mcw.edu/phenotypes/

[9]http://pga.mcw.edu/

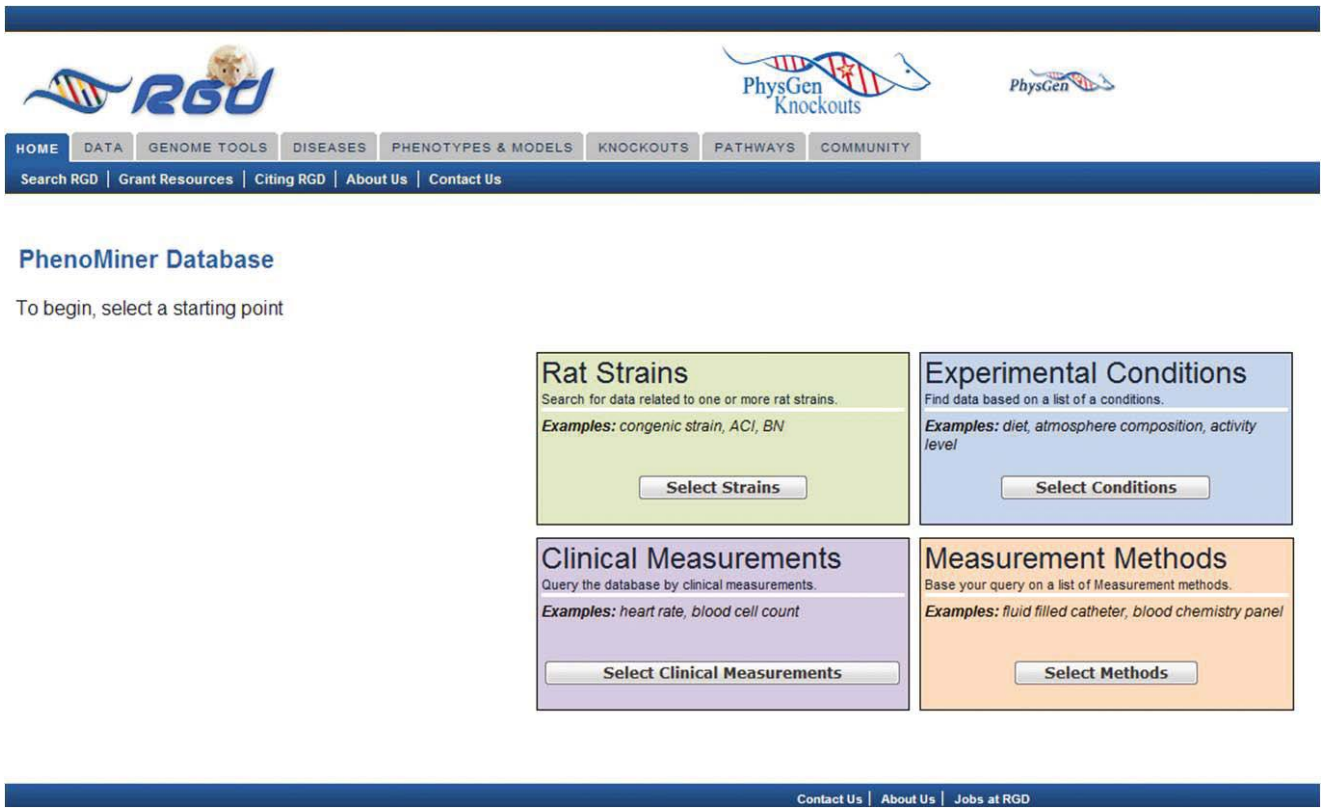


FIGURE 6 | The PhenoMiner website.

and organ weights have been conducted under baseline conditions for inbred and mutant strains.

Because the rat is an ideal model organism for pharmacology, biochemistry, and physiology research, the published literature is a rich resource of data on phenotype measurements for particular strains. Papers reporting cardiovascular, respiratory, renal, morphological, and blood chemistry measurement data as well as those with measurements related to cancer were targeted for the initial phase of the PhenoMiner resource. Over 13,000 measurement records from these three sources have been mapped to the three ontologies and integrated into PhenoMiner (**Figure 7**).

DISCUSSION AND CONCLUSION

The three ontologies created have proven to be excellent tools for standardizing phenotype measurement data for projects involving a wide variety of data types and data sources. Targeting the three basic elements of: (1) what was measured; (2) how it was measured; and (3) under what conditions it was measured, facilitated standardization while allowing for flexibility in providing associated information such as units of measurement or duration of condition to be formatted in ways particular to the integrating resource. These ontologies allowed phenotype measurement data from disparate studies to be integrated without compromising study-specific aspects related to methodology. Multiple datasets of human epidemiological data and rat phenotype data were successfully integrated into resources designed to meet the needs of diverse research communities even though the underlying technological framework for the databases and associated tools differed. While integrating varied phenotype datasets was the primary motivation for the development of these ontologies, they can be deployed in a variety of other projects as well. Because of their availability at NCBO, they can be utilized with the NCBO Annotator, a Web service that annotates journal abstracts[10] which facilitates curation efforts and queries for appropriate literature for specific projects. Because of their focus on experimental data, the use of these ontologies in text mining tools would also help investigators identify and prioritize literature.

Creating structures to integrate phenotype measurement data from multiple sources is an important task as investigators draw on the strength of the genomic and sequence variation resources to identify underlying genotype factors related to phenotypes and diseases. In order to make these connections, researchers need to easily access and analyze phenotype measurement data related to individuals and various model strains, and information on experimental conditions and methodologies that may affect the measurement values. Employing multiple ontologies to standardize data formats facilitates the integration of these vital datasets and provides the structure on which innovative data mining, analysis, and presentation tools can be built. These types of resources can provide researchers with a more accurate

[10]http://bioportal.bioontology.org/annotator#

Study ID	Sample ID	CMO	CMO value	MMO	MMO site	XCO 1	XCO 1 Dur	XCO 1 Value	XCO 1 Ord	XCO 2	XCO 2 Dur	XCO 2 Value	XCO 2 Ord
2	1	Heart rate	264.6 beats/ min	Isolated perfused heart, balloon tipped catheter		Air oxygen content	14 days	12%	1	Controlled sodium content diet	14 days	0.4%	1
2	2	Heart rate	247.2 beats/ min	Isolated perfused heart, balloon tipped catheter		Air oxygen content	14 days	12%	1	Controlled sodium content diet	14 days	0.4%	1
2	3	Heart rate	260.7 beats/ min	Isolated perfused heart, balloon tipped catheter		Air oxygen content	14 days	21%	1	Controlled sodium content diet	14 days	0.4%	1
6	1	Heart rate	271 beats/ min	Fluid filled catheter	Femoral artery	Air oxygen content		21%	1				
6	2	Heart rate	239.7 beats/ min	Fluid filled catheter	Femoral artery	Air carbon dioxide content	7 mins	7%	1				
5	1	Heart rate	470.7 beats/ min	Fluid filled catheter	Femoral artery	Walking on treadmill	3 min	0.8 m/min	1				
5	2	Heart rate	457.8 beats/ min	Fluid filled catheter	Femoral artery	Walking on treadmill	3 min	0.8 m/min	1	Running on treadmill	3 min	1.6 m/min	2

FIGURE 7 | Example of phenotype measurement data from multiple studies mapped to the three ontologies for clinical measurement, measurement method, and experimental condition.

picture of phenotype variations among populations and as well as the impact of measurement methods may have on measurement results. The influence of experimental and environmental conditions on phenotypes and disease will also be easier to elucidate when researchers have access to large numbers of measurements from a wide variety of studies. This is an important step in helping investigators link genotypes to phenotypes. Finally, the use of multiple ontologies to standardize data elements into single quantifiable records can be used in many paradigms to integrate datasets. Convergence among phenotyping efforts can be fostered using this methodology through the use of existing ontologies, such as MP, PATO, and HPO, in conjunction with ontologies that further refine the phenotype data. Engaging other communities will provide the platform for data mining across species or across phenotyping programs using a variety of phenotyping protocols.

AUTHORS' CONTRIBUTIONS

Mary Shimoyama created the ontologies, organized the data mapping, and integration at the Medical College of Wisconsin and participated in data mapping for the COVER project. Rajni Nigam assisted in ontology term creation and participated in data mapping. Leslie Sanders McIntosh organized data mapping and informatics component creation for the COVER project. Rakesh Nagarajan participated in the informatics infrastructure design for COVER project. Treva Rice coordinated data mapping and integration for the COVER project. D. C. Rao participated in design and coordinated data mapping and integration for the COVER project. Melinda R. Dwinell participated in the design of the PhenoMiner tool and data integration at the Medical College of Wisconsin. All authors have read and approved the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of the RGD curation team at the Medical College of Wisconsin and the staff and informatics team at Washington University. This project is funded in part by R01HL094271, R01HL094286, and R01HL064541.

REFERENCES

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's online Mendelian inheritance in man (OMIM). *Nucleic Acids Res.* 37, D793–D796.

Bouchard, C., Leon, A. S., Rao, D. C., Skinner, J. S., Wilmore, J. H., and Gagnon, J. (1995). The HERITAGE family study. Aims, design, and measurement protocol. *Med. Sci. Sports Exerc.* 27, 721–729.

Butte, A. J., and Kohane, I. S. (2006). Creation and implications of a phenome-genome network. *Nat. Biotechnol.* 24, 55–62.

Day-Richter, J., Harris, M. A., Haendel, M., and Lewis, S. (2007). OBO-Edit – an ontology editor for biologists. *Bioinformatics* 23, 2198–2200.

Degtyarenko, K., Hastings, J., De Matos, P., and Ennis, M. (2009). ChEBI: an open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics* Chapter 14, Unit 14.19.

Gkoutos, G. V., Green, E. C., Mallon, A. M., Hancock, J. M., and Davidson, D. (2004). Building mouse phenotype ontologies. *Pac. Symp. Biocomput.* 178–189.

Goujon, M., Mcwilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* 38(Suppl.), W695–W699.

Gu, D., He, J., Hixson, J. E., Jaquish, C. E., Liu, D., Rao, D. C., Whelton, P. K., Yao, Z., He, J., Bazzano, L. A., Chen, C.-S., Chen, J., Hamm, L., Muntner, P., Reynolds, K., Reuben, J. R., Whelton, P. K., Yang, W., Rao, D. C., Brown, M., Gu, C., Rice, T., Schwander, K., Wang, S., Gu, D., Cao, J., Chen, J., Duan, X., Huang, J., Huang, J., Li, J., Liu, D., Liu, D., Pan, E., Wei, Y., Wu, X., Lu, F., Jin, S., Meng, Q., Wu, F., Zhao, Y., Ma, J., Li, W., Zhang, J., Hu, D., Ding, Y., Wen, H., Zhang, M., Zhang, W., Ji, X., Li, R., Zu, H., Yao, C., Li, Y., Shen, C., Zhou, J., Mu, J., Chen, E., Huang, Q., Wang, M., Yao, Z.-J., Chen, S., Gu, D., Li, H., Wang, L., Zhang, P., Zhao, Q., Hixson, J. E., Shimmin, L. C., and Jaquish, C. E. (2007). GenSalt: rationale, design, methods and baseline characteristics of study participants. *J. Hum. Hypertens.* 21, 639–646.

Hancock, J. M., Adams, N. C., Aidinis, V., Blake, A., Bogue, M., Brown, S. D., Chesler, E. J., Davidson, D., Duran, C., Eppig, J. T., Gailus-Durner, V., Gates, H., Gkoutos, G. V., Greenaway, S., Hrabe De Angelis, M., Kollias, G., Leblanc, S., Lee, K., Lengger, C., Maier, H., Mallon, A. M., Masuya, H., Melvin, D. G., Muller, W., Parkinson, H., Proctor, G., Reuveni, E., Schofield, P., Shukla, A., Smith, C., Toyoda, T., Vasseur, L., Wakana, S., Walling, A., White, J., Wood, J., and Zouberakis, M. (2007). Mouse Phenotype Database Integration Consortium: integration [corrected] of mouse phenome data resources. *Mamm. Genome* 18, 157–163.

Kaminuma, E., Mashima, J., Kodama, Y., Gojobori, T., Ogasawara, O., Okubo, K., Takagi, T., and Nakamura, Y. (2010). DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res.* 38, D33–D38.

Kwitek, A. E., Jacob, H. J., Baker, J. E., Dwinell, M. R., Forster, H. V., Greene, A. S., Kunert, M. P., Lombard, J. H., Mattson, D. L., Pritchard, K. A. Jr., Roman, R. J., Tonellato, P. J., and Cowley, A. W. Jr. (2006). BN phenome: detailed characterization of the cardiovascular, renal, and pulmonary systems of the sequenced rat. *Physiol. Genomics* 25, 303–313.

Morgan, H., Beck, T., Blake, A., Gates, H., Adams, N., Debouzy, G., Leblanc, S., Lengger, C., Maier, H., Melvin, D., Meziane, H., Richardson, D., Wells, S., White, J., Wood, J., De Angelis, M. H., Brown, S. D., Hancock, J. M., and Mallon, A. M. (2010). EuroPhenome: a repository for high-throughput mouse phenotyping data. *Nucleic Acids Res.* 38, D577–D585.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biol.* 11, R2.

Robinson, P. N., and Mundlos, S. (2010). The human phenotype ontology. *Clin. Genet.* 77, 525–534.

Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., Sim, I., Chute, C. G., Solbrig, H., Storey, M. A., Smith, B., Day-Richter, J., Noy, N. F., and Musen, M. A. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS* 10, 185–198.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., John Wilbur, W., Yaschenko, E., and Ye, J. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 38, D5–D16.

Serikawa, T., Mashimo, T., Takizawa, A., Okajima, R., Maedomari, N., Kumafuji, K., Tagami, F., Neoda, Y., Otsuki, M., Nakanishi, S., Yamasaki, K., Voigt, B., and Kuramoto, T. (2009). National BioResource Project-Rat and related activities. *Exp. Anim.* 58, 333–341.

Shimoyama, M., Petri, V., Pasko, D., Bromberg, S., Wu, W., Chen, J., Nenasheva, N., Kwitek, A., Twigger, S., and Jacob, H. (2005). Using multiple ontologies to integrate complex biological data. *Comp. Funct. Genomics* 6, 373–378.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.

Smith, C. L., and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 1, 390–399.

Smith, C. L., Goldsmith, C. A., and Eppig, J. T. (2005). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* 6, R7.

Sorsby, A. (1965). Gregor Mendel. *Br. Med. J.* 1, 333–338.

Williams, R. R., Rao, D. C., Ellison, R. C., Arnett, D. K., Heiss, G., Oberman, A., Eckfeldt, J. H., Leppert, M. F., Province, M. A., Mockrin, S. C., and Hunt, S. C. (2000). NHLBI family blood pressure program: methodology and recruitment in the HyperGEN network. Hypertension genetic epidemiology network. *Ann. Epidemiol.* 10, 389–400.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 February 2012; paper pending published: 05 March 2012; accepted: 30 April 2012; published online: 28 May 2012.

Citation: Shimoyama M, Nigam R, McIntosh LS, Nagarajan R, Rice T, Rao DC and Dwinell MR (2012) Three ontologies to define phenotype measurement data. Front. Gene. ***3****:87. doi: 10.3389/fgene.2012.00087*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

METHODS ARTICLE
published: 22 June 2012
doi: 10.3389/fgene.2012.00111

Development and use of ontologies inside the neuroscience information framework: a practical approach

Fahim T. Imam*, Stephen D. Larson, Anita Bandrowski, Jeffery S. Grethe, Amarnath Gupta and Maryann E. Martone*

Neuroscience Information Framework, Center for Research in Biological Systems, University of California San Diego, La Jolla, CA, USA

Edited by:
John Hancock, Medical Research Council, UK

Reviewed by:
Douglas M. Bowden, University of Washington School of Medicine, USA
Qiangfeng Cliff Zhang, Columbia University, USA

****Correspondence:***
Fahim T. Imam and Maryann E. Martone, Neuroscience Information Framework, Center for Research in Biological Systems, University of California San Diego, La Jolla, CA 92093-0446, USA.
e-mail: mimam@ucsd.edu; memartone@ucsd.edu

An initiative of the NIH Blueprint for neuroscience research, the Neuroscience Information Framework (NIF) project advances neuroscience by enabling discovery and access to public research data and tools worldwide through an open source, semantically enhanced search portal. One of the critical components for the overall NIF system, the NIF Standardized Ontologies (NIFSTD), provides an extensive collection of standard neuroscience concepts along with their synonyms and relationships. The knowledge models defined in the NIFSTD ontologies enable an effective concept-based search over heterogeneous types of web-accessible information entities in NIF's production system. NIFSTD covers major domains in neuroscience, including diseases, brain anatomy, cell types, sub-cellular anatomy, small molecules, techniques, and resource descriptors. Since the first production release in 2008, NIF has grown significantly in content and functionality, particularly with respect to the ontologies and ontology-based services that drive the NIF system. We present here on the structure, design principles, community engagement, and the current state of NIFSTD ontologies.

Keywords: ontologies, ontology reuse, neuroscience ontology, semantic search

INTRODUCTION

The Neuroscience Information Framework Project (NIF)[1] facilitates the utilization of the growing number of neuroscience-relevant data available through the web. NIF, supported by the National Institutes of Health Blueprint, was initiated in recognition of the current difficulties of locating and searching across the diverse array of web-based resources and databases (Gardner et al., 2008). The NIF was also charged with developing tools and strategies for creating resources that can be integrated across neuroscience domains. The end product is a semantic search engine and a knowledge discovery portal that consists of a framework for describing neuroscience resources and provides simultaneous access to multiple types of information organized by relevant categories. Through its extensive resource catalog and data federation, NIF currently represents the largest source of neuroscience information available on the web.

The semantic framework through which these diverse resources are accessed is provided by the NIF Standardized Ontologies (NIFSTD; Bug et al., 2008). NIFSTD represents an extensive collection of terms and concepts from the major domains of neuroscience. The overall ontology has been assembled in a form that promotes reuse of multiple existing biomedical ontologies and standard vocabulary sources, while allowing for extension and modification over the course of its evolution. This paper presents the development principles of NIFSTD along with its application within the NIF system.

[1]NIF, http://neuinfo.org

NIFSTD DESIGN PRINCIPLES

As originally proposed in Bug et al. (2008), NIFSTD was envisioned as an extensive set of ontologies, specific to the domain of neuroscience. NIFSTD started its journey with a carefully designed set of principles which enabled its ontologies to be maximally reusable, extendable, and practically applicable within information systems. Over the course of its evolution, NIFSTD augmented its principles in order to conform to the current, up-to-date trends, and practices recommended by the semantic web communities as well as by the community of standard biomedical ontologies. NIFSTD closely follows the OBO Foundry (Smith et al., 2007) best practices; however, the constraints of the NIF project required that we take a practical approach, designed to easily extend the NIFSTD ontologies, while at the same time mitigating against any disruptions to the production NIF system. Our approach is outlined following the discussion of the NeuroLex Semantic Wiki framework in Section "The NeuroLex Semantic Wiki Framework."

NIFSTD MODULAR STRUCTURE

The NIFSTD ontologies are built in a modular fashion, where each module covers a distinct, orthogonal domain of neuroscience (Bug et al., 2008). Modules covered in NIFSTD include anatomy, cell types, experimental techniques, nervous system function, small molecules, and so forth. The upper-level classes in NIFSTD modules are carefully normalized under the classes of Basic Formal Ontology (BFO)[2]. These normalizations closely follow the guidelines specified in BFO manual (BFO manual)[3]. Based on the

[2]BFO, http://www.ifomis.org/bfo
[3]BFO manual, http://www.ifomis.org/bfo/manual

principles described in Rector (2003), NIFSTD utilizes a powerful ontology modularization technique that allows its ontologies to be reusable and easily extendable. Each domain specified in **Table 1** has their corresponding module in NIFSTD. The individual module in turn may cover multiple sub-domains. The ingestion strategy for each source in **Table 1** is shown in the "Import/Adapt" column, where "import" refers to the BFO compliant sources which were already represented in OWL; "adapt" refers to the sources that required refactoring of the source vocabularies into OWL, and/or required normalization under BFO entities.

NIFSTD REPRESENTATION FORMALISM

NIFSTD modules are expressed in W3C standard Web Ontology Language (OWL)[4]; Description Logic (OWL-DL) formalism. Using OWL-DL, NIFSTD provides a balance between its expressivity and computational decidability. OWL-DL also allows the NIFSTD ontologies to be supported by a range of open source DIG compliant reasoners (DIG Group)[5] such as Pellet and Fact++. NIFSTD utilizes these reasoners to maintain its inferred classification hierarchies as well as to keep its ontologies in a logically consistent state.

NIFSTD currently supports OWL 2 (OWL 2 Primer)[6], the latest ontology language advocated by the W3C consortium. OWL 2 provides improved ontological features such as defining property chain rules to enable transitivity across object properties, specifying reflexivity, asymmetry, and disjointness between object properties, richer data-types, qualified cardinality restrictions, and enhanced annotation capabilities.

ACCESSING NIFSTD ONTOLOGIES

NIFSTD is available in OWL format[7] for loading in Protégé (Protégé Ontology Editor)[8] or other ontology editing tools that use the OWL API. Protégé has been the main editing tool for building the NIFSTD modules. Currently, NIFSTD supports Protégé 4.X versions with OWL 2. On the web, NIFSTD is available through the NCBO BioPortal (NIFSTD in NCBO BioPortal)[9], which also provides annotation and various mapping services. NIFSTD is also available in RDF and has its SPARQL endpoint (NIFSTD SPARQL endpoint)[10].

Within NIF, NIFSTD is served through an ontology management system called OntoQuest (Gupta et al., 2008, 2010). Originally reported in Chen et al. (2006), OntoQuest generates an OWL-compliant relational schema for NIFSTD ontologies and implements various graph search algorithms for navigating, path finding, hierarchy exploration, and term searching in ontological

[4]OWL, http://www.w3.org/TR/owl-ref/
[5]DIG Group, http://dl.kr.org/dig/
[6]OWL 2 Primer, http://www.w3.org/TR/owl2-primer/
[7]OWL format, http://purl.org/nif/ontology/nif.owl
[8]Protégé Ontology Editor, http://protege.stanford.edu/
[9]NIFSTD in NCBO BioPortal, http://bioportal.bioontology.org/ontologies/40510
[10]NIFSTD SPARQL endpoint, http://ontology.neuinfo.org/sparql-endpoint.html

Table 1 | The NIFSTD OWL modules and corresponding community sources from which they were built.

NIFSTD modules	External source	Import/adapt
Organismal taxonomy	NCBI Taxonomy, GBIF, ITIS, IMSR, Jackson Labs mouse catalog; the model organisms in common use by neuroscientists are extracted from NCBI taxonomy and kept in a separate module with mappings	Adapt
Molecules, chemicals	IUPHAR ion channels and receptors, sequence ontology (SO); NIDA drug lists from ChEBI, and imported protein ontology (PRO)	Adapt/import
Sub-cellular anatomy	Sub-cellular anatomy ontology (SAO). Extracted cell parts and sub-cellular structures from SAO-CORE. Imported GO cellular component with mapping	Adapt/import
Cell	CCDB, NeuronDB, NeuroMorpho.org. Terminologies; OBO cell ontology was not considered as it did not contain region specific cell types	Adapt
Gross anatomy	NeuroNames extended by including terms from BIRNLex, SumsDB, BrainMap.org, etc.; multi-scale representation of nervous system, macroscopic anatomy	Adapt
Nervous system function	Sensory, behavior, cognition terms from NIF, BIRN, BrainMap.org, MeSH, and UMLS	Adapt
Nervous system dysfunction	Nervous system disease from MeSH, NINDS terminology; Imported Disease Ontology (DO) with mapping	Adapt/import
Phenotypic qualities	Phenotypic quality ontology (PATO); imported as part of the OBO foundry core	Import
Investigation: reagents	Overlaps with molecules above from ChEBI, SO, and PRO	Adapt/import
Investigation: instruments, protocols, plans	Based on the ontology for biomedical investigation (OBI) to include entities for biomaterial transformations, assays, data collection, data transformations. OBI-Proxi class still remains. See discussion below	Adapt
Investigation: resource type	NIF, OBI, NITRC, biomedical resource ontology (BRO)	Adapt
Investigation: cognitive paradigm	Cognitive paradigm ontology (CogPO) was extended from NIF-investigation module	Import
Biological process	Gene ontology (GO) biological process	Import

This table reports the updates of the external sources that were previously used in Bug et al. (2008) paper.

graphs. OntoQuest provides a collection of web services to extract specific ontological content[11]. Ontoquest also provides the NIF search portal with automated query expansion (Gupta et al., 2010) for matching NIFSTD terms, including those that are defined through logical restrictions.

REUSE OF EXTERNAL SOURCES

One of the founding principles of NIFSTD is to avoid duplication of efforts by conforming to existing standard biomedical ontologies and vocabulary sources. It should also be noted that NIF is not charged with developing new ontological modules but relies on community sources for new contents. Whenever possible, NIFSTD reuses those existing sources as the initial building blocks for its core modules. Essentially, these external sources were selected based on their relevance to neuroscience knowledge models. **Table 1** illustrates the modules in NIFSTD that are either adapted, or imported, or extracted from external community sources. NIFSTD reuses a diverse collection of sources for its ontologies. These sources range from fully structured ontologies to loosely structured controlled vocabularies, lexicons, or nomenclatures that exist within the biomedical community. Each module in NIFSTD (**Table 1**) integrates the relevant terms or concepts from those external sources into a single, internally consistent ontology with a matching standard nomenclature. The process and nature of reusing an external source in NIFSTD varied upon its state. The following rules summarize the basic reuse principles:

1. If the source is already represented in OWL, normalized under BFO, and is orthogonal to existing NIFSTD modules, the source is simply imported as a new module.
2. If the source is represented in OWL and orthogonal to NIFSTD modules, but is not normalized under BFO, then an ontology-bridging module (explained later) is constructed before importing the new source. These kinds of bridging modules declare the necessary relational properties to normalize the target ontology source under BFO.
3. If the source is orthogonal to NIFSTD modules, but is not represented in OWL, or does not use BFO as its foundational layer, then the source should be converted into OWL, and should be normalized under BFO following the Second rule above.
4. If the source is satisfiable by the above three principles but observed to be too large for NIF's scope, then a relevant subset is extracted as suggested by NIF domain experts.

For the ontologies that are of type 4 above, NIFSTD currently follows MIREOT principles (Courtot et al., 2009) that allow extracting a required subset of classes from a large ontology, e.g., ChEBI, NCBI Organismal Taxonomy, etc.

Neuroscience Information Framework Project readily accepts contributions from groups working on ontologies in the neuroscience domain. For example, the Cognitive Paradigm Ontology (CogPO; Turner and Laird, 2012), has been imported under the NIF-Investigation module. As we worked through the process of adopting CogPO, we needed to make sure that the upper-level classes in CogPO were BFO compliant and derivable under the same foundational layers of NIFSTD, and the properties were extended from OBO-RO. As part of NIFSTD, CogPO can be used to annotate datasets for specific querying and comparisons and the contents are exposed via NeuroLex for community involvement (see The NeuroLex Semantic Wiki Framework).

At the beginning of the NIF project, the size, format, or immaturity of some community ontologies necessitated that NIF add significant custom content in order to provide coverage in certain modules. Over the last couple of years, the tools for extracting relevant portions of ontologies and for converting ontologies from OBO to OWL format have been improved. Thus, since the last publication (Bug et al., 2008), several of these custom ontologies were swapped for community ontologies. However, it should be noted that the NIF-Investigation module still contains "OBI-proxy" classes that were originally meant to be replaced by the matured version of OBI under BFO 1.0. However, the matured version of OBI entailed many of the original OBI-proxy classes to be retired, changed their identifiers, and sometimes did not replace them by any new classes. As NIF-Investigation continued to add many new concepts under the original obi-proxy classes, directly importing the current OBI to replace the proxy classes was not a reasonable solution. However, we have proposed the NIF-Investigation terms to be added, aligned, and maintained within OBI. We plan to incorporate portions of OBI to be extracted under NIF-Investigation, for the future release of NIFSTD.

SINGLE INHERITANCE FOR NAMED CLASSES

An asserted named class in NIFSTD can have only one named class as its parent. However, the same named class can be asserted under multiple anonymous classes. This principle promotes the named classes to be univocal to avoid ambiguities. In NIFSTD, classes with multiple parents are derivable via automated classification on defined classes. This approach saves a great deal of manual labor and minimizes human errors inherent in maintaining multiple hierarchies. Also, this approach provides logical and intuitive reasons as to how a class may exist under multiple, different hierarchies. A useful example can be seen in Neuronal type classification in section "Example Knowledge Model: NIFSTD Neuronal Cell Types" where a particular neuron type can be a subclass of multiple different "anonymous" classes, e.g., Neuron X is a Neuron that has GABA as a neurotransmitter. The details about the motivation behind this approach can be found in Alan Rector's Normalization pattern discussion (Ontology Design Pattern: Normalization)[12].

UNIQUE IDENTIFIERS AND ANNOTATION PROPERTIES

NIFSTD entities are named by unique identifiers and are accompanied by a variety of annotation properties. These annotation properties are mostly derived from Dublin Core Metadata (DC) and the Simple Knowledge Organization System (SKOS) model. While several annotation properties still exist from the legacy modules of BIRNLex, from which NIFSTD was built (Bug et al., 2008), currently NIFSTD only requires the following set of annotation properties for a given new class.

[11] OntoQuest, http://ontology.neuinfo.org/ontoquest-service.html

[12] Ontology Design Pattern: Normalization, http://ontologydesignpatterns.org/wiki/Submissions:Normalization

- rdfs: label – A human-readable name for a class or property. If a class can be named in multiple ways, a label is chosen based on the name most commonly used in literatures as selected by NIF domain experts. Other names for the class can be kept as synonyms.
- nifstd: createdDate – The date when the current class or property was created. This property serves as a way to track versioning.
- dc: contributor – Name of the curator who has contributed to the definition of a class.
- core: definition – A natural language definition of a class. In ideal case, this definition should be written in a standard Aristotelian form.
- nifstd: definitionSource – A traceable source for the current definition in a free text form. A source could be a URI, an informal publication reference, a PubMed ID, etc.
- owl: versionInfo – A version number associated with NeuroLex category.

The following set of properties is used when necessary:

- nifstd: modifiedDate – The date when the current class was last updated.
- nifstd: synonym – A lexical variant of the class name.
- nifstd: abbreviation – A short name serving as a synonym, consisting of a sequence of letters typically taken from the beginning of words of which either the preferred label or another synonym are composed. Note that this should only be used for standard abbreviation (i.e., those that are commonly used in literatures, e.g., in a PubMed indexed article)[13]. Many of the abbreviations supplied are actually acronyms, but we no longer distinguish between the two.
- rdfs: comment – Anything related to the class or the property that should be noted.

For the current versions of Protégé, the above properties can be set as the default set of properties for NIFSTD. NIFSTD has other annotation properties associated with version control which will be described in Section "Versioning policy." When extracting external sources using MIREOT principles, NIFSTD keeps the identical source URIs along with the original identifier fragments unaltered. This approach allows NIF to avoid extra mapping efforts with the community sources. Prior to the MIREOT approach, the practice was simply to assign new class ID for any externally sourced classes which led to maintenance difficulty due to too many mapping annotations. We still have some mappings from the BIRNLex vocabularies, as we did not have the MIREOT tool when we started.

NIFSTD OBJECT PROPERTIES

NIFSTD imports the OBO Relations Ontology (OBO-RO) for the standard set of properties as defined by the OBO Biomedical community. Other object properties in NIFSTD are mostly derived from OBO-RO. Based on where the relations are asserted, there are two kinds of relations that exist in NIFSTD: one that are within a same module, i.e., intra-modular relations, and the other that is inter-modular, cross-domain relations that exist as a separate, isolated module between two independent modules.

The intra-modular relations are the ones that exist as universally true within the classes of a specific module; these relations are kept integrated together within the same module. The relations between entities that could vary based on a specific application and require domain-dependent viewpoints are kept in a separate bridging module – a module that only contains logical restrictions and definitions on a required set of classes assigned between multiple modules (see **Figure 1**).

The bridging modules allow the core domain modules – e.g., anatomy, cell type, etc., to remain independent of one another. This approach keeps the modularity principles intact, and facilitates broader communities to utilize and extend NIFSTD with reasonable ease. Some of the bridge modules in NIFSTD are constructed in order to include simple semantic equivalencies between ontologies.

New bridging modules can be developed should a user desire a customized ontology of their own application domain based on one or multiple NIFSTD core modules. For example, the Neurodegenerative Disease Phenotype Ontology (NDPO; Maynard et al., submitted) is essentially a bridge module that asserts a number of entity-quality relations (on classes in relevant NIFSTD modules) to specify and define a list of named phenotypes.

As the existing reasoners fail to scale against large ontologies like NIFSTD, modularity in NIFSTD plays an important role. From an ontology development perspective, it is crucial to frequently check the consistency after asserting any new set of classification along with their axioms. Since NIFSTD is divided into smaller independent modules, the task of automated classification and consistency checking becomes much more maintainable while working on a specific module of interest.

VERSIONING POLICY

NIFSTD provides various levels of versioning for its content. It allows humans and machine to choose the level of version information required for tracking changes. Various annotation properties are associated with versioning different levels of content, including creation and modified date for each of the classes and files, file level versioning for each of the modules, and annotations for retiring antiquated concept definitions, tracking former ontology graph position, and replacement concepts.

– NIFSTD: has Former Parent Class – the full logical URI of the former parent class of a deprecated class or any other class whose super-class has been changed. This property is typically used for a deprecated/retired class.
– NIFSTD: is Replaced By Class – the full logical URI of the new class that exists as the replacement of the current retired class. This property should only be used if there exist a new replacing class.

The umbrella file nif.owl at http://purl.org/nif/ontology/nif.owl always imports the current versions of the NIFSTD modules. All other versions after the 1.0 release can be accessed from the NIF ontology archive at http://ontology.neuinfo.org/NIF/Archive/.

[13]PubMed, http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0006431/

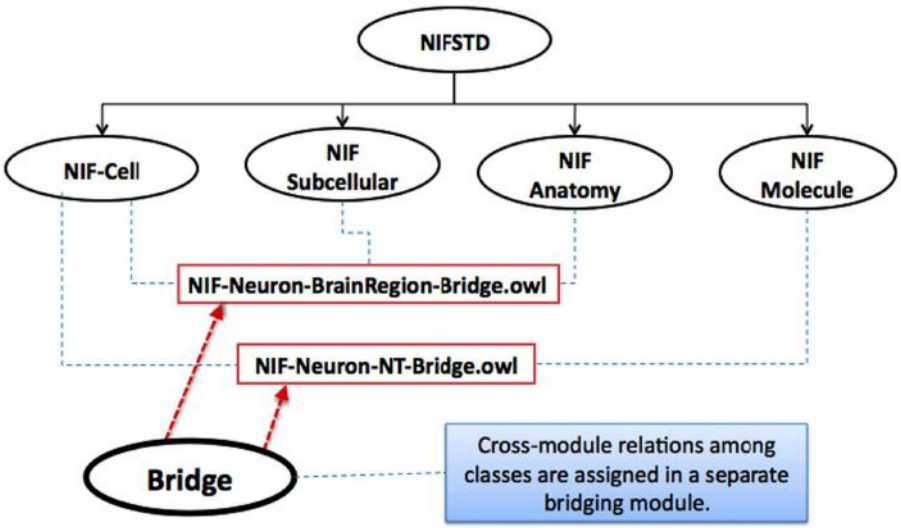


FIGURE 1 | Two example bridging OWL modules in NIFSTD (rectangular boxes) that contain class property associations between multiple core modules.

THE NeuroLex SEMANTIC WIKI FRAMEWORK

One of the largest roadblocks that NIF identified early in the project was the lack of tools for domain experts to view, edit, and contribute their knowledge to the formal ontologies like NIFSTD. When constructing its ontologies, NIF strived to balance the involvement of the neuroscience community for domain expertise and the knowledge engineering community for ontology expertise. By combining several open sourced, semantic media wiki technologies, NIF created NeuroLex, a semantic wiki for the neuroscience community and domain experts. Details about the NeuroLex platform will be included in a separate publication (Larson et al., in preparation). Here we focus on the interplay between the NeuroLex and NIFSTD.

RELATION BETWEEN NIFSTD AND NeuroLex

The initial contents of the NeuroLex were derived from NIFSTD which established its neuroscience-centric semantic framework and enabled the semantic relationships among its category pages. NIFSTD OWL classes were automatically transformed into category pages containing simplified, human-readable class descriptions. The category pages are editable and readily available to access, annotate, or enhance by the community or domain experts. Additions of new categories and enhancements to the NeuroLex contents are regularly transformed into NIFTSD in formal OWL-DL expressions. NeuroLex category pages are linked with NIF Search interface where users can quickly view descriptive ontological details about a matching search term.

While the properties in NeuroLex are meant for easier interpretation, the corresponding restrictions in NIFSTD are more rigorous and based on standard OBO-RO relations. For example, the property "soma located in" is translated as "Neuron X" has_part some ["Soma" and (part_of some "Brain region Y")] in NIFSTD. Sometimes similar kinds of "macro" relations, e.g., "has_neurotransmitter," are used in NIFSTD, recognizing that these relations can be defined in a more rigorous manner if required. These macro relations can be defined as a composition of multiple transitive properties using OWL 2.0 property chains.

Neuroscience Information Framework Project considers NeuroLex.org as the main entry point for the broader community to access, annotate, edit, and enhance the core NIFSTD content. The peer-reviewed contributions in the media wiki are later implemented in formal OWL modules. As NIF relies on the communities to enhance its ontologies, NeuroLex is an ideal interface for NIF's current scope. For example, it has proven to be effective in the area of neuronal cell types where NIF is working with a group of neuroscientists to create a extensive list of neurons and their properties.

NIFSTD/NeuroLex CURATION WORKFLOW

The NIFSTD development/curation workflow includes the tasks mentioned in each of the boxes followed by a number as in **Figure 2**. Each of the steps along with the associated tasks in the workflow is summarized in the following table, **Table 2**.

THE SCOPE OF NeuroLex

NeuroLex can be viewed as a full-fledged information management system that provides a bottom-up ontology development approach where multiple participants can edit the ontology instantly. The semantics of NeuroLex are limited to what is convenient for the domain experts. Essentially, the NeuroLex approach is not a replacement for top-down ontology construction, but critical to increase accessibility for non-ontologist domain experts. NeuroLex provides various simple forms for structured knowledge where communities can contribute and verify their knowledge with ease. It also allows the simple query mechanisms to generate specific class hierarchies, or extraction of a specific portion of the ontology contents based on certain properties in a spreadsheet, without having to learn any complicated ontology tools.

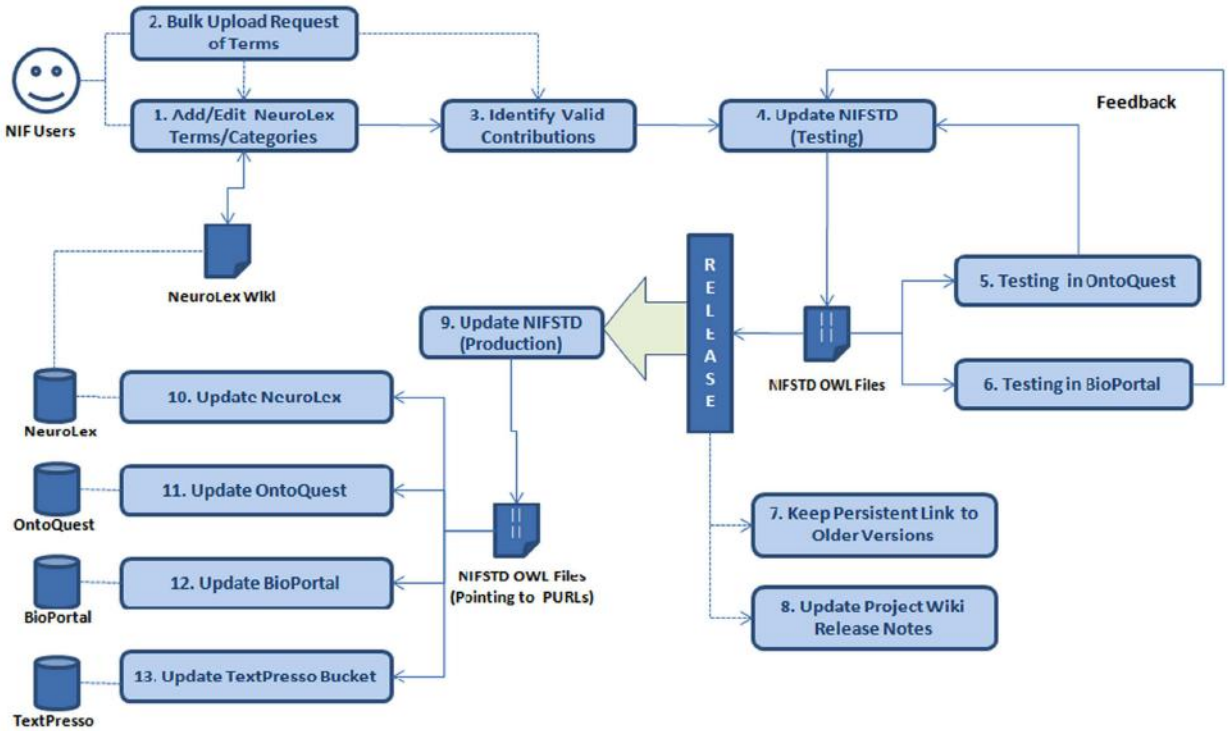


FIGURE 2 | Transition of contributions between the NeuroLex and the NIFSTD.

Table 2 | The steps and tasks involved in NIFSTD/NeuroLex curation workflow.

Step	Tasks
Add/edit to NeuroLex	This step involves various NIF users/group who are interested in adding, updating, enhancing, or annotating the vocabularies through the NeuroLex wiki
Bulk upload	Depending on the number and nature of terms (i.e., adding new large sub-tree of an existing class, or new classes with known parents for a specific NIF module, etc.), we can support bulk upload of terms
Identify valid contributions	This step involves identifying the contributions in the previous steps that are valid according to a NIF domain expert. This step should make sure that a term contributed is actually new and not a synonym or duplicate of any existing term. Invalid contributions should be rolled back in NeuroLex during this step
Update NIFSTD (testing)	This step involves updating the NIFSTD OWL files or creating new OWL files in testing environment based on the update of contents from previous steps
Testing in OntoQuest	After significant updates in NIFSTD (every 1–1.5 months), the OWL implementations should be loaded in OntoQuest testing server for feedbacks
Testing in BioPortal	After significant updates in NIFSTD (every 1–1.5 months), the OWL implementations should be tested in BioPortal staging environment for feedbacks
Persist links to older versions	After positive feedbacks from Step 5 and 6, we archive the links to the old OWL files and post the links to the project wiki
Release notes	Before releasing the production version of NIFSTD, a new release note should be added for the forthcoming version. The release note should include a version number, version specific major changes, major hierarchical changes, newly added module(s), and other technical changes
Update NIFSTD (production)	This step involves updating the NIFSTD OWL modules in the production server that are pointing to the Persistent URLs (PURL; http://purl.org)
Update NeuroLex Wiki	A new release of NIFSTD should be followed by the updates in NeuroLex wiki which will reflect the implemented additions/changes of the NIFSTD contents merged with the previous iteration
Update OntoQuest	A major release of NIFSTD should be followed by an update in OntoQuest production version
Update BioPortal	A major release of NIFSTD should be followed by an update in BioPortal production version

Although NeuroLex does not support many of the standard first-order logic features that are available in standard OWL-DL formalism to support reasoning, we feel that NeuroLex has its place within the process of standard ontology development. NeuroLex can be seen as an interface to initiate the process of conceptualization where the main target is to associate the categories/concepts with the existing set of concepts/categories using simple properties. Users contributing to the NeuroLex are not formal knowledge engineers, but domain scientists tasked with ensuring that the appropriate concepts and relationships are available to the NIF for effective search and description of NIF resources.

Essentially, NeuroLex is a place to accommodate the concepts and entities that are found in literatures and other legitimate sources that are not yet been realized within a formal ontology relevant to Neuroscience. NeuroLex allows a neuroscientist to add a new concept without having to worry about its deep semantic consequence due to incompleteness or partial truth about an asserted. Fundamentally, OWL-DL can only represent a conceptual domain in a rigorous, logical fashion where it can only reason over a set of statements that are asserted to be true. Unlike OWL-DL version in NIFSTD, incomplete, non-rigorous knowledge is fine within the context of NeuroLex. Over time a concept/category in Neurolex can become ideally matured in a collaborative and completely transparent manner. As the conceptual model becomes more mature in NeuroLex, the category pages become more interconnected. While transitioning these NeuroLex contents into NIFSTD, the fundamental idea is to identify and append all the necessary logical constraints on top those "interconnection" properties. The transition of knowledge from NeuroLex to NIFSTD is essentially a context-aware, "structured" transition of knowledge between a group of domain experts and formal oncologists. This, in fact, is a practical approach of developing life science ontologies in a collaborative manner.

NeuroLex VS. WIKIPEDIA

Although both NeuroLex Wiki and Wikipedia projects share some common goals of providing a platform for collaborative knowledge development, they differ significantly in terms of their available functionalities, features, and scopes. In order to expose structured knowledge, WikiPedia utilizes MediaWiki templates through its "info-boxes." These info-boxes are transformed into RDF graphs by the DBPedia project in order to mine the knowledge structures. Building on top of Semantic MediaWiki (an extension of Mediawiki platform), NeuroLex does not require the two step process of producing the RDF knowledge models. Unlike Wikipedia, where a user must learn the wiki-text syntax to contribute her knowledge, NeuroLex provides "Semantic Forms" option for easy editing. NeuroLex contributors therefore can choose not to be confronted with wiki-text syntax for editing.

Figure 3 illustrates some of the unique features of the NeuroLex wiki platform. A standard Wikipedia page requires all the knowledge about the page to be entered manually within a single text box. In contrast, as NeuroLex has a semantic backend to structure its overall knowledge, a page in NeuroLex can dynamically call relevant information from other pages. For example, NeuroLex has the ability to automatically assemble related knowledge about Cerebellum as shown in the boxes corresponding to **Figures 3D–F**. Note that the information contained in **Figures 3D–F** are not entered as

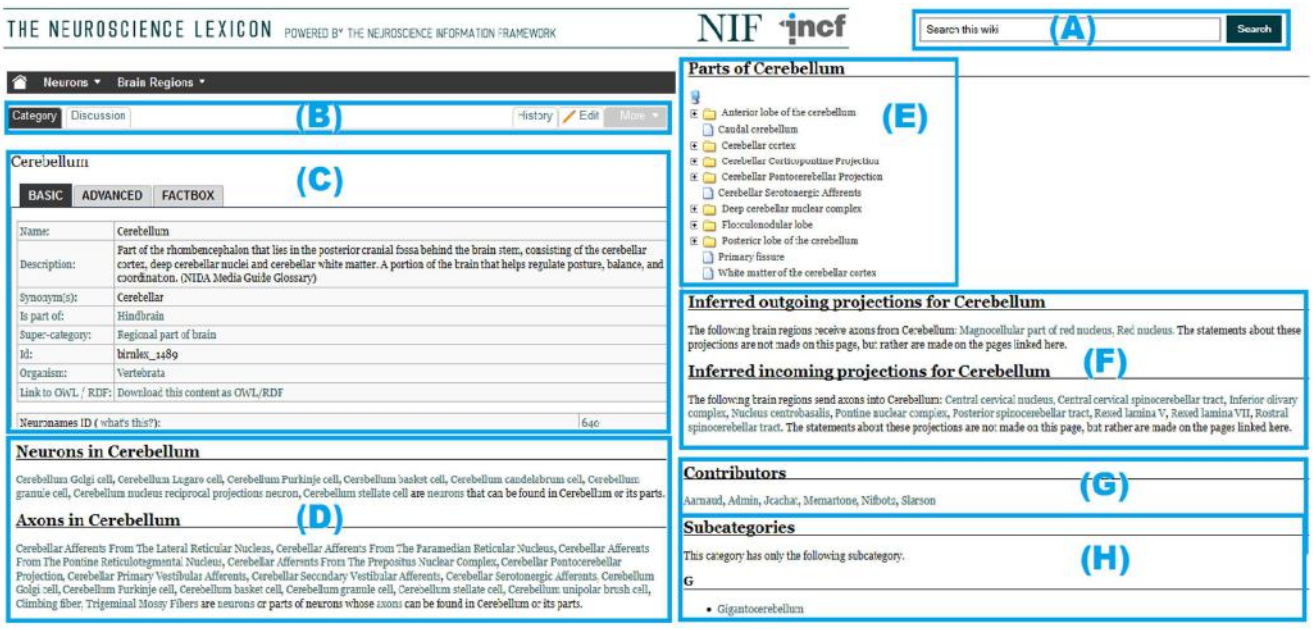


FIGURE 3 | Structure of contents in a typical NeuroLex category page. (A) The standard input text field for searching the entire NeuroLex wiki contents. **(B)** Different tabs to display and edit the contents of a particular category page. **(C)** The structured contents of a category page (e.g., Cerebellum). Boxes corresponding to **(D–F)** demonstrate the ability of the NeuroLex to automatically assemble related knowledge about a particular category from the edits made in other NeuroLex pages. **(G)** The list of contributors who made edits to the page. **(H)** The list of subcategories of a particular category page.

part of the "Cerebellum" page itself, but are automatically assembled from the edits made to other pages, e.g., if a user enters a soma location for a neuron that is a part of cerebellum, the neuron automatically shows up on this page under the "Neurons in cerebellum" in **Figure 3D**. Analogously, the "Axons in Cerebellum" in **Figure 3D** is also populated from the edits made in other pages. Finally, NeuroLex is meant to house all concepts of relevance to neuroscience, regardless of whether or not they are particularly noteworthy.

EXAMPLE KNOWLEDGE MODEL: NIFSTD NEURONAL CELL TYPES

Following the basic NIFSTD principle, NIF neuron types are listed in a simple, flat hierarchy of named classes under the common super-class called "Neuron" within the NIF-Cell module. These cell types were largely contributed by the NIF team, as the Cell Ontology (CL) did not contain many region specific cell types (Bard et al., 2005) at the time NIF-cell was developed. The neurons in NIFSTD are asserted with logical necessary conditions based on a set of properties that characterize mature neurons and provide a reasonable basis on which to classify them. The relational properties relate neuron types in NIF-Cell module with classes in other modules such as NIF-Subcell, NIF-Anatomy, NIF-Quality, and NIF-Molecule. As mentioned earlier in section "NIFSTD Design Principles," these cross-module relations are kept in separate bridging modules. These modules contain necessary restrictions along with a set of defined classes to infer useful classification of neurons. The following list illustrates some of the key neuron types along with their classification schemes:

- Neurons by their soma location in different brain regions – e.g., Hippocampal neuron, Cerebellum neuron, Retinal neuron
- Neurons by their neurotransmitter – e.g., GABAergic neuron, Glutamatergic neuron, Cholinergic neuron
- Neurons by their circuit roles – e.g., Intrinsic neuron, Principal neuron
- Neurons by their morphology – e.g., Spiny neuron
- Neurons by their molecular constituents – e.g., Parvalbumin neuron, Calretinin neuron.

One of the most powerful features of having an ontology is that it allows explicit knowledge of a domain to be asserted from which implicit logical consequences can be inferred using logical reasoners. The following example illustrates the strength and usefulness of this feature. NIFSTD includes various neuron types with an asserted simple hierarchy under the common superclass, "Neuron." **Figure 4** illustrates an example with five neuron types.

However, as illustrated in **Figure 5**, logical restrictions about these neurons are asserted in a bridging module along with a set of defined neuron types with necessary and sufficient conditions. The first table in **Figure 5** defines three neuron types with logical necessary and sufficient conditions: the Cerebellum neuron, Principal neuron, and GABAergic neuron. The second table in **Figure 5** lists a set of necessary restrictions for Cerebellum Purkinje cell. All these restrictions written in a readable format here are expressed in OWL-DL in actual NIFSTD. When the NIF-Cell module along with the bridging modules are passed to a reasoner, the reasoner automatically computes for the asserted neuron types and produces a hierarchy where the neurons are inferred under multiple superclasses. In this example, although the Cerebellum Purkinje cell was not asserted under any specific named neuron types, after invoking the automated reasoner, the neuron becomes an inferred subclass of four different defined neurons – namely, the GABAergic neuron, Cerebellum neuron, Spiny neuron, and Principal neuron as illustrated in **Figure 6**.

Note that NIF does not currently perform deep logical modeling of neuron types, such that a reasoner would be able to deduce the necessary and sufficient conditions for a neuron to be considered a Purkinje cell. It is currently very difficult to provide universal identifying criteria for identification of particular cell types (Hamilton et al., 2012). Rather, NIF uses the logical restrictions placed on properties to generate useful classifications of neurons based on general properties that can be used to enhance search within the NIF portal, and which allows neurons to be grouped based on common features. As the ontologies are also available in RDF graphs, SPARQL queries can be written to extract a list of data elements that are linked through these simple properties.

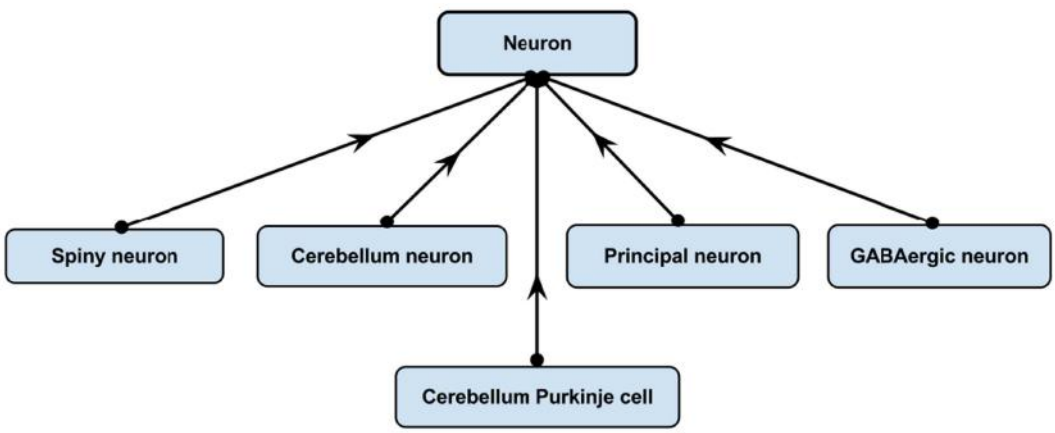


FIGURE 4 | Asserted simple hierarchy of "Cerebellum Purkinje cell."

Class Name	Defining Expression (Necessary & Sufficient Conditions)
Cerebellum neuron	**Is a 'Neuron'** whose soma lies in any part of the **'Cerebellum'**.
Principal neuron	**Is a 'Neuron'** which has **'Projection neuron role'**; i.e., any neuron whose axon projects out of the brain region in which its soma lies.
GABAergic neuron	**Is a 'Neuron'** that has **'GABA'** as a neurotransmitter.
Spiny neuron	**Is a 'Neuron'** which has **'Spiny dendrite quality'**.

Class Name	Asserted Restrictions (Necessary Conditions)
Cerebellum Purkinje cell	1. Is a **'Neuron'**. 2. Its soma lies within the **'Purkinje cell layer of cerebellar cortex'**. 4. It has **'Projection neuron role'**. 5. It has **'GABA'** as a neurotransmitter. 6. It has **'Spiny dendrite quality'**.

FIGURE 5 | Typical NIFSTD restrictions asserted for various neuron types.

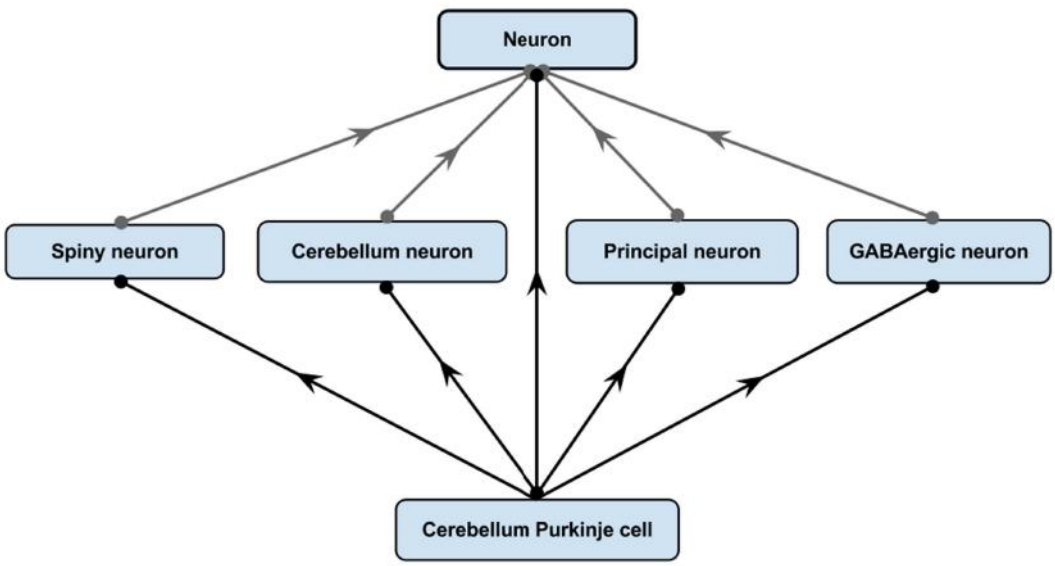


FIGURE 6 | After invoking a reasoner NIFSTD Cerebellum Purkinje cell becomes a subclass of four different defined neuron types based on the restrictions specified in Figure 5.

EVOLUTION OF NIFSTD

Since the first release in 2008, the NIFSTD ontologies have undergone extensive revision and refinements. These updates include simplified structural changes to its import hierarchies, retirement of duplicate classes due to multiple imports from the first release, enforced modularization principles by adopting bridging modules between the core modules, enhancement into the partonomy restrictions in NIF Gross Anatomy, refactoring the modules under more appropriate BFO classes, simplifying the NIFSTD backend module that comprises the common entities shared by all of the NIFSTD modules. As biomedical ontologies from different communities matured, NIFSTD included various new modules such as the Gene Ontology (GO), Protein Ontology (PRO), part of ChEBI, and Human Disease Ontology (DOID). NIFSTD also imported a simplified, slim version of NCBI Taxonomy removing taxon ranks not commonly used by neuroscientists (Gardner et al., 2008). Various equivalency bridge modules have been constructed in order to ensure logical mappings on the overlapping classes between the existing NIFSTD modules and newly added modules. NIFSTD core contents have also been rapidly enhanced from NeuroLex contributions. The vision that was proposed in 2008 (Bug et al., 2008) of building detailed representations of multi-scale

brain structure using common and interconnected building blocks has been realized in NIFSTD v1.8 and subsequent versions, as illustrated above with NIFSTD's representation of neuronal cell types.

An example of how the NIFSTD continues to evolve is shown by the NIFSTD gross anatomy module. While constructing the original gross anatomy module, NIF avoided importing Foundational Model of Anatomy (FMA) or Mouse Anatomy as we wanted the core module to represent generic, species independent parts. NIFSTD extensively adopted and transformed portions of NeuroNames (Bowden et al., 2012) structures into an OWL ontology to represent NIF's brain anatomy without any species-specific restrictions. Initially, NIFSTD divided up the brain parts into several categorical superclasses. These different categorical classes were established to make it easier to keep different types of brain parts straight, without having to worry too much about assigning other relations. These super categories included the following parts:

- Regional part: A division of a structure that can be recognized by gross anatomical features, cytoarchitecture or chemoarchitecture, e.g., cerebral cortex is a regional part of brain.
- Cytoarchitectural part: A division of a brain structure that is based on the organization of cell bodies, usually revealed by a Nissl stain, e.g., CA1 is a cytoarchitectural part of the hippocampus.
- Chemoarchitectural part: A division of a brain structure based on the distribution of some chemical marker, e.g., the patch/matrix division of the caudate nucleus
- Aggregate part: A brain structure that is composed of many different parts that are distributed in location, e.g., basal ganglia.
- Composite part spanning many brain regions: A brain part whose subdivisions are found throughout the neuraxis, e.g., the corticospinal tract.

For the current version of NIFSTD, these categorical classes are removed from the primary hierarchy of the brain structures, as they have been largely replaced through the assignment of "part of" relationships. NIF currently considers all parts of brain as a "regional part of brain" at the highest level to represent a general reference structure across species. Through the partonomy restrictions, parts comprising groupings of brain structures such as white matter structures, basal ganglia, and circumventricular organs can be generated, so that they can be used in the NIF search system. A more detailed report on the representation of brain parts within NIFSTD, in conjunction with the program on ontologies of the International Neuroinformatics Coordinating Facility[14] is in preparation.

USE OF NIFSTD WITHIN THE NIF SYSTEM

As outlined in the introduction, the NIFSTD provides the semantic framework for searching across the diverse data sources available through the NIF. As such, it was designed to represent high level neuroscience knowledge that is useful for searching data sources. The NIF portal provides simultaneous search across three major sources of information: (1) The NIF Registry; a catalog of >4500 resources (databases, tools, materials, services) categorized according to the NIF Resource module and annotated with keywords derived from other NIFSTD modules; (2) The NIF Data Federation: Deep access to the contents of >150 databases; and (3) NIF Literature: Abstracts of Pub Med and full text of open access articles.

Neuroscience Information Framework Project adopted a very aggressive population strategy to ensure that the system was well populated as rapidly as possible in order to serve its primary mission of providing deep access to neuroscience-relevant data and tools. As is well known, resources are developed with little thought to how they would interoperate within a global information system, leading to a fragmented system of custom resources, each with their own data models and terminologies. Just as with the NIFSTD itself, we designed the system to be able to work with resources in their current state, while building in capacity for us to evolve the system over time, as new tools and technologies became available. The NIFSTD is not meant to represent the information within these sources; rather, it serves as a semantic index for searching across those diverse resources. In other words, the semantic search mechanism in NIF is enhanced through the utilization of NIFSTD; as the ontology becomes richer, search is improved. Through OntoQuest, NIF enhances the search by providing an ontology-based query formulation, source selection, term expansion, and finally better ranking on the search results based on the NIFSTD contents.

Using OntoQuest services, search through the NIF interface auto-completes to terms within the NIFSTD. OntoQuest provides automatic expansion of these terms to their synonyms, abbreviations, and lexical variants as defined in NIFSTD. The NIF system uses a query language inspired by current search engines like Google. In this language, the simplest option is to ask a keyword query, but one can optionally add predicates on metadata and data attributes, specify return structures, and make references to ontologies. An advanced search box allows users to expand terms into their ontologically related terms, e.g., part of, subclasses that can be included within the search. NIF employs Boolean operators to connect these terms in an intelligent fashion, i.e., all synonyms are joined through an "OR" operator as are any related classes selected via the ontology tree. Additional concepts entered into the search box are joined through an "AND." Thus, if a user enters "Neurodegenerative disease""drug," and selects Parkinson's disease and Alzheimer's disease as children of neurodegenerative disease, NIF will join them as follows (synonyms are omitted: "Neurodegenerative disease OR Parkinson's disease OR Alzheimer's disease" AND "drug"). Typical query expansion constructs are presented in **Table 3** illustrating how the contents from ontologies are utilized.

One of the key features of the current NIFSTD is the inclusion and enrichment of various cross-domain bridging modules which include a number of useful defined classes. As illustrated in the neuronal examples in section "Evolution of NIFSTD," we have been working with domain experts to define relationships between entities within different NIFSTD core modules, e.g., brain region to neuron; neuron to molecule that weave together the different modules in a coherent manner. These defined classes are then used by the NIF system to formulate its useful concept-based queries through OntoQuest. For example, while searching for "GABAergic

[14]International Neuroinformatics Coordinating Facility, http://incf.org

Table 3 | Examples of ontological query expansions in NIF through OntoQuest.

Example query type	Ontological expansion
A single term query for hippocampus and its synonyms	`synonyms(Hippocampus);` expands to Hippocampus OR "Cornu ammonis" OR "Ammon's horn" OR "hippocampus proper"
A conjunctive query with three terms	`transcription AND gene AND pathway`
A sixth term and/or query with one term expanded into synonyms	`(gene) AND (pathway) AND (regulation OR "biological regulation") AND (transcription) AND (recombinant)`
A conjunctive query with two terms, where a user chooses to select the subclasses of the second term	`synonyms(zebrafish AND descendants(promoter,subclassOf)))`, zebrafish gets expanded by synonym search and the second term transitively expands to all subclasses of promoter as well as their synonyms
A single term query for an anatomical structure where a user chooses to select all of the anatomical parts of the term along with synonyms	`synonyms(descendants(Hippocampus,partOf))`, expands to all parts of hippocampus and all their synonyms through the ontology. All parts are joined as an "OR" operation
A conjunctive query with two terms, where a user chooses to select all the equivalent terms for the second term	`synonyms(Hippocampus) AND equivalent(synonyms(memory))`, the second term uses the ontology to find all terms that are equivalent to the term memory by ontological assertion, along with synonyms
A conjunctive query with two terms, where a user is interested in a specific subclasses for both of the terms	`synonyms(x:descendants(neuron,subclassOf) where x.neurotransmitter=`"GABA") AND synonyms(gene where gene. name="IGF"), x is an internal variable
A query to seek all subclasses of neuron whose soma location is in any transitive part of the hippocampus	`synonyms(x:descendants(neuron,subclassOf) where x.soma.location` = descendants (Hippocampus, partOf))
A query to seek a conceptual term that is semantically equivalent to a collection of terms rather than a single term	`"GABAergic neuron" AND equivalent ("GABAergic neuron")`, The term gets recognized as ontologically *equivalent* to any neuron that has GABA as a neurotransmitter and therefore expands to a list of inferred neuron types

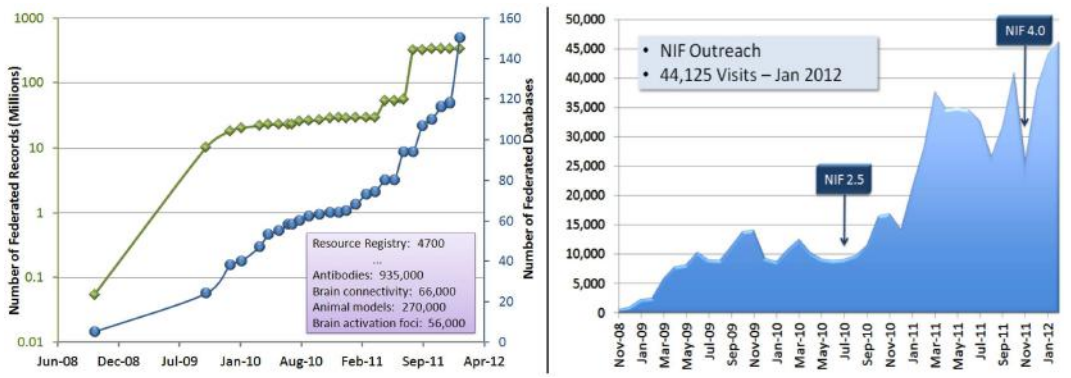


FIGURE 7 | On the left, the increase of NIF contents in terms of the number of federated records (green) and databases (blue). On the right, the increase of community outreach in terms of the number of visitors to the NIF portal.

neuron," the NIF query expansion through OntoQuest recognizes the term as "defined" from the ontology, and looks for any neuron that has GABA as a neurotransmitter (instead of the lexical match of the search term) and enhances the query over those inferred list of neurons. Searching this defined concept in a Google search would essentially exclude all the GABAergic neurons unless they are explicitly listed within the search box. Other analogous example include query formulation for the defined concepts like Tracer, Anterograde tracer, Retrograde tracer, Neurotransmitter, Neurotransmitter receptor, Non-human primate, Drug of abuse, etc.

Since the first release in 2008, NIF has grown significantly in contents and community building. The chart on the left in **Figure 7** illustrates the growth of federated records and database resources in NIF since June, 2008. The chart on the right illustrates the utilization growth in visits per month across NIF holdings, including NIF search portal, NeuroLex, and NIF services. Currently, NIF search portal has ∼6000 visits per month, and NeuroLex has over 15,000 visits per month. Also, it is worth mentioning that a significant number of current NIF users are successfully finding their desired terms and concepts from the NIFSTD vocabularies. For example, based on the recent Google analytics report (from April 1st to 30th, 2012) on NIF's user interaction patterns, out of total 7108 search events, 3317 committed auto-complete search (i.e., 46.66% of the desired search terms existed in NIFSTD vocabularies), and 256 of them required advanced ontological query expansion search.

CONCLUSION

The NIF project provides an example of practical ontology development and how it can be used to enhance search and data integration across diverse resources. NIF uses the NIFSTD to provide a semantic index to heterogeneous data sources and the basis of the concept-based query system. Using the upper-level BFO ontologies allowed us to promote a broad semantic interoperability between a large numbers of biomedical ontologies. The modularity principles along with the bridging modules allowed us to limit the complexity of the base ontologies. Users of NIFSTD can exclude the NIF specific bridging modules, which promotes easy extendibility and keeps the modularity principles intact. All of the practices adopted by NIF were designed to allow ontologies to be utilized within an evolving production system with minimum disruption as the ontologies and ontology design principles evolved.

We have defined a process to form complex semantics to various neuroscience concepts through NIFSTD and through NeuroLex collaborative environment. NIF encourages the use of community ontologies for resource providers, and as the project moves forward, we are using NIFSTD to build an increasingly rich knowledge base for neuroscience that integrates the data sources with the larger life science community. Essentially, the key aspects of these knowledge-bases are the integration of necessary semantic layer on top of the data elements found in databases, and literature corpus by linking those data elements with ontological concepts. NIF is closely following the movements such as Open Data, Linked Data, and Web of Data, to provide effective new ways that could semantically integrate data regardless of their sources.

ACKNOWLEDGMENTS

Supported for NIF is provided by a contract from the NIH Neuroscience Blueprint HHSN271200800035C via the National Institute on Drug Abuse.

REFERENCES

Bard, J., Rhee, S. Y., and Ashburner, M. (2005). An ontology for cell types. *Genome Biol.* 6, R21.

Bowden, D. M., Song, E., Kosheleva, J., and Dubach, M. F. (2012). NeuroNames: an ontology for the BrainInfo portal to neuroscience on the web. *Neuroinformatics* 10, 97–114.

Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., Larson, S. D., Rubin, D., Shepherd, G. M., Turner, J. A., and Martone, M. E. (2008). The NIFSTD and BIRNLex vocabularies: building extensive ontologies for neuroscience. *Neuroinformatics* 6, 175–194.

Chen, L., Martone, M. E., Gupta, A., Fong, L., and Wong-Barnum, M. (2006). "Ontoquest: exploring ontological data made easy," in *Proceedings 31st International Conference on Very Large Database (VLDB)*, Seoul, 1183–1186.

Courtot, M., Gibson, F., Lister, A., Malone, J., Schober, D., Brinkman, R., and Ruttenberg, A. (2009). *MIREOT: The Minimum Information to Reference an External Ontology Term.* Available at: http://dx.doi.org/10.1038/npre.2009.3576.1

Gardner, D., Goldberg, D. H., Grafstein, B., Robert, A., and Gardner, E. P. (2008). Terminology for neuroscience data discovery: multi-tree syntax and investigator-derived semantics. *Neuroinformatics* 6, 161–174.

Gupta, A., Bug, W. J., Marenco, L., Condit, C., Rangarajan, A., Müller, H. M., Miller, P. L., Sanders, B., Grethe, J. S., Astakhov, V., Shepherd, G., Sternberg, P. W., and Martone, M. E. (2008). Federated access to heterogeneous information resources in the neuroscience information framework (NIF). *Neuroinformatics* 6, 205–217.

Gupta, A., Condit, C., and Qian, X. (2010). BioDB, an ontology-enhanced information system for heterogeneous biological information. *Data Knowl. Eng.* 69, 1084–1102.

Hamilton, D. J., Shepherd, G. M., Martone, M. E., and Ascoli, G. A. (2012). An ontological approach to describing neurons and their relationships. *Front. Neuroinformatics* 6:15. doi:10.3389/fninf.2012.00015

Rector, A. (2003). "Modularisation of domain ontologies implemented in description logics and related formalisms including OWL," in *Proceedings of K-CAP 2003*, Sanibel Island.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.

Turner, J. A., and Laird, A. R. (2012). The cognitive paradigm ontology: design and application. *Neuroinformatics* 10, 57–66.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 23 February 2012; accepted: 29 May 2012; published online: 22 June 2012.

Citation: Imam FT, Larson SD, Bandrowski A, Grethe JS, Gupta A and Martone ME (2012) Development and use of ontologies inside the neuroscience information framework: a practical approach. Front. Gene. ***3****:111. doi: 10.3389/fgene.2012.00111*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

OPINION ARTICLE
published: 26 November 2012
doi: 10.3389/fgene.2012.00269

An ontological analysis of some biological ontologies

Briti Deb*†

Bioinformatics Center, IISc, Bangalore, India
**Correspondence: briti@ieee.org*

Edited by:
John Hancock, University of Cambridge, UK

Reviewed by:
John Hancock, University of Cambridge, UK

INTRODUCTION

The functional importance of biological entities makes their understanding, analysis, and representation essential in modern biology. Arguably, semantic representation necessary for machine interoperability is a far more difficult task than syntactic representation, necessitating conceptual schema and ontologies for in-silico biological knowledge representation. Biological ontologies are increasingly being developed for prediction, big data integration in semantic web, visualization, unstructured data interpretation, annotation, and eHealth ontology. Despite being widely used, deficiencies exist (Kumar and Smith, 2003; Kumar et al., 2004; Mougin and Bodenreider, 2005; Pal, 2006; Schulz, 2006) in their concepts, relations, and frameworks in general, leading to difficulties in semantic interoperability and integration, and possibility of wrong prediction after using them. In this opinion article, I attempted for the first time (in my knowledge) to show that some characteristic inadequacies of biological ontologies could be detected and prevented by using the philosophically inspired OntoClean method (Guarino, 2002) and the top-level DOLCE ontology (Masolo et al., 2009), both of which have well-founded formal semantics, and finally proposed an outline of a novel ontology framework which aims to remove existing deficiencies. Though preliminary, my arguments suggest that it would be worthy to look deeper into the use of OntoClean and DOLCE toward detecting ontological inadequacies and improving them, a detailed analysis of which is left as a future work. I may state that, this discussion is not meant to criticize any of the ontologies, but to present some arguments on their respective design choices when seen in the light of OntoClean and DOLCE.

ANALYSIS WITH OntoClean AND DOLCE

The OntoClean method proposes to tag concepts on a taxonomy according to the following philosophical meta-properties: rigid, anti-rigid, non-rigid, carry-identity-criterion, supply-identity-criterion, carry-unity, and carry-anti-unity. It must be noted here that, these assignments are not "definitive" (Guarino and Welty, 2004), rather it demonstrate logical consequences of making such choices. In the following, I present six cases and put forward my conjectures on detecting ontological inadequacies and solutions to correct them using the OntoClean method and DOLCE top-level ontology.

(a) OntoClean method suggests that, an entity has an essential property if that property is held by it all the time, and is rigid if all the instances possess that property (Guarino, 1998, 1999). Adult human beings would have an essential property of "adult behavior." But due to the fact that Gene Ontology (GO) (Ashburner et al., 2000) terms are designed to be applied across many species, a term such as the "adult behavior" could lead to confusion when applied to unicellular organisms like amoeba. It could also be debated whether the GO term "adult behavior" is a rigid property or not, since all instances of human adults may not display adult behavior. I believe that modeling ontologies after considering essential and rigid properties of entity would prevent such an inadequacy.

(b) Identity criteria is used to recognize whether individual entities are the same or different (Guarino, 1998, 1999). Several characteristic inadequacies both in the GO and the Unified Medical Language System (UMLS) could be identified (Mougin and Bodenreider, 2005), as a result of the failure to draw distinction between continuant (i.e., endurant) and occurrent (i.e., perdurant) entities (Masolo et al., 2009), and between dependent (such as cellular motion, temperature, and mass) and independent entities (Kumar and Smith, 2003). In the UMLS, a function is a continuant which has a subsumption relation with a process (an occurant), which I believe could be a case of identity violation. Instead of using the subsumption (is_A) relation, using the "participate_In" relation such as, "A Continuant participate_In an Occurant" would bring in more ontological adequacy.

(c) The GO described the term "extracellular" as the space external to the outermost structure of a cell. A question could arise on deciding the location and/or the granularity level of the term extracellular (Kumar et al., 2004). This problem could be attributed to the fact that the GO has not explicitly modeled the identity criteria of entities such as the extracellular, to be able to recognize entities as the same or different entity, in addition to not recognizing the unity criterion necessary toward recognizing parts of these individual entities.

(d) According to the UMLS, an organism attributes is_A conceptual entity. Given the fact that, organism attribute

†The author is currently a graduate student at the Institute of Computer Science, Tartu 50409, Estonia.

is not necessarily dependent on mind (because all organisms need not have a mind), whereas a conceptual entity is necessarily dependent on mind, my conjecture is that identity criteria has been violated. Using the DOLCE top-level ontological distinctions, and reorganizing conceptual entity as an agentive-physical-object (DOLCE:APO) and organism attribute as a non-agentive-physical-object (DOLCE:NAPO) could have helped to detect such inconsistencies.

(e) In the GO, the term "GO:0020037:heme binding" is a molecular function. From (Guarino, 1999, 2002), I understand that material role is a role which is anti-rigid (−R), inherit identity (+I), and dependent (+D). I believe that this GO term could be well modeled as a material role, having OntoClean meta-properties such as (−R, +I, +D), and it could be subsumed by the type called "molecular function," resulting to more semantic clarity. In the BFO, role has been subsumed by dependent entity which is subsumed by continuant entity (Kumar and Smith, 2003). Placing role under "property" which isA DOLCE:Universal, rather than assuming role enduring self-identically through time as is in BFO (Kumar and Smith, 2003) seems to me as a better choice.

(f) In the Open Biomedical Ontologies (OBO) (http://obo.sourceforge.net), relations lack explicit formal definitions creating the possibility of confusions. Inadequacies could also be found in the use of relations such as is_A and part_Of (Smith et al., 2005; Burek et al., 2006). The distinction between function and their functioning in the GO has also been confusing, though a solution was attempted by the GO by appending the term "activity," e.g., "galactokinase activity" (Krummenacker et al., 2009). Another problem which could arise from the use of multiple inheritances and is_A overloading is polysemy (Guarino, 1999). The problem of multiple inheritance in its conceptual hierarchies prevents it from logical reasoning applications. To understand one such inadequacy, let's take an example from the GO described graphically in Krummenacker et al. (2009). If galactokinase activity is made a subclass of carbohydrate kinase activity and phosphotransferase activity, then as per the rules of subsumption (Guarino, 1998), it would inherit the identity of both the super-classes. But, I believe this creates confusion, since the identity criteria of carbohydrate kinase activity would be different from the identity criteria of phosphotransferase activity, and any prediction based on such a hierarchy could lead to erroneous results. Though it may also appear as a semantic duplication in the ontology, the reasons why I feel it is important are: (1) lack of maintainability, (2) increased chances of confusion/inconsistency, (3) reduced search time efficiency, and (4) extra storage space. The formal logical modeling techniques in OntoClean method and top-level ontological distinctions between "universal" and "particular" in DOLCE, both having well founded formal semantics, could be used to understand better the underlying ontological structure and semantics of the classes and avoid polysemy.

CONCLUSIONS AND FUTURE DIRECTIONS

Biological ontologies are plagued by deficiencies in conceptual integration and inter-linkage (Beisswanger et al., 2007), and lacks sufficient concepts to represent functioning/actions/events (Schulz, 2006). The primary aim of this paper is to argue for the use of DOLCE (supported by OntoClean methods) as an upper level (or foundational) ontology, to describe general concepts shared by several biological domain ontologies, and to align them. As a semantic web agent may use several domain ontologies, aligning the domain ontologies becomes crucial to reduce semantic mismatch among services. Arguably, mathematical knowledge, comprising both symbolic notations and natural language, remains largely under-represented for semantic web agents. Though MathML and OpenMath have been developed to be used with Resource Description Framework (RDF), their success have been limited by the vocabulary provided by the ontology. As DOLCE (and OntoClean) have not been used so far as a foundational ontology for aligning many widely used biological domain ontologies, this discussion is intended as a motivation for a more detailed future research on it. As an example of how DOLCE could capture ontological categories underlying mathematical knowledge, the parthood relation in DOLCE could be used to represent: "a symbol is part_Of a formula." Complementarity of foundational ontology and domain ontologies is believed to serve as a corrective to each others individual pitfalls. Detailed analysis of how DOLCE can satisfy all the requirements to represent mathematical knowledge is left as a future work.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Beisswanger, E., Schulz, S., Stenzhorn, H., and Hahn, U. (2007). BIOTOP: an upper domain ontology for the life sciences. *Appl. Ontol.* 3, 205–212.

Burek, P., Hoehndorf, R., Loebe, F., Visagie, J., Herre, H., and Kelso, J. (2006). A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics* 22, e66–e73.

Guarino, N. (1998). "Some ontological principles for designing upper level lexical resources," in *Proceedings of the First International Conference on Language Resources and Evalution* (Granada), 28–30.

Guarino, N. (1999). "The role of identity conditions in ontology design," in *Proceedings* of IJCAI-99 *W*orkshop on Ontologies and Problem-Solving Methods (Stockholm).

Guarino, N. (2002). Evaluating ontological decisions with OntoClean. *Commun. ACM* 45, 61–65.

Guarino, N., and Welty, C. A. (2004). "An overview of OntoClean," in *The Handbook on Ontologies*, eds S. Staab and R. Studer (Berlin: Springer-Verlag), 151–172.

Krummenacker, M., Siegele, D. A., Hu, J. C., Karp, P. D., and Keseler, I. M. (2009). What we can learn about *Escherichia coli* through application of Gene Ontology. *Trends Microbiol.* 17, 269–278.

Kumar, A., Smith, B., and Novotny, D. D. (2004). Biomedical informatics and granularity. *Comp. Func. Genomics* 5, 501–508.

Kumar, A., and Smith, B. (2003). *The Unified Medical Language System and the Gene Ontology: Some Critical Reflections, KI 2003: Advances in Artificial Intelligence.* eds R. Kruse, A. Günter, and B. Neumann (Berlin: Springer Verlag), 135–148.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2009).

WonderWeb Deliverable D18, Ontology Library (final). IST Project 2001-33052. WonderWeb: Ontology Infrastructure for the Semantic Web.

Mougin, F., and Bodenreider, O. (2005). Approaches to eliminating cycles in the UMLS Metathesaurus: naïve vs. formal. *AMIA Annu. Symp. Proc.* 2005, 550–554.

Pal, D. (2006). On gene ontology and function annotation. *Bioinformation* 1, 97–98.

Schulz, S. (2006). Towards an upper-level ontology for molecular biology. *AMIA Annu. Symp. Proc.* 2006, 694–698.

Schulz, S., Beisswanger, E., Hahn, U., Wermter, J., Kumar, A., and Stenzhorn, H. (2006). "From GENIA to BioTop towards a toplevel ontology for biology," in *The 4th International Conference on Formal Ontology in Information Systems* (Baltimore), 103–114.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., and Lomax, J. (2005). Relations in biomedical ontologies. *Genome Biol.* 6, R46.

Received: 21 October 2012; accepted: 06 November 2012; published online: 26 November 2012.

*Citation: Deb B (2012) An ontological analysis of some biological ontologies. Front. Gene. **3**:269. doi: 10.3389/fgene.2012.00269*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 28 June 2012
doi: 10.3389/fgene.2012.00115

The choice between MapMan and Gene Ontology for automated gene function prediction in plant science

Sebastian Klie*[1] *and Zoran Nikoloski*[2]*

[1] *Genes and Small Molecules Group, Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany*
[2] *Systems Biology and Mathematical Modeling Group, Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany*

Edited by:
John Hancock, Medical Research Council, UK

Reviewed by:
John Hancock, Medical Research Council, UK
Pankaj Jaiswal, Oregon State University, USA
Keiichi Mochida, RIKEN, Japan

****Correspondence:***
Zoran Nikoloski, Systems Biology and Mathematical Modeling Group, Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm D-14476, Germany.
e-mail: nikoloski@mpimp-golm.mpg.de

Since the introduction of the Gene Ontology (GO), the analysis of high-throughput data has become tightly coupled with the use of ontologies to establish associations between knowledge and data in an automated fashion. Ontologies provide a systematic description of knowledge by a controlled vocabulary of defined structure in which ontological concepts are connected by pre-defined relationships. In plant science, MapMan and GO offer two alternatives for ontology-driven analyses. Unlike GO, initially developed to characterize microbial systems, MapMan was specifically designed to cover plant-specific pathways and processes. While the dependencies between concepts in MapMan are modeled as a tree, in GO these are captured in a directed acyclic graph. Therefore, the difference in ontologies may cause discrepancies in data reduction, visualization, and hypothesis generation. Here provide the first systematic comparative analysis of GO and MapMan for the case of the model plant species *Arabidopsis thaliana* (*Arabidopsis*) with respect to their structural properties and difference in distributions of information content. In addition, we investigate the effect of the two ontologies on the specificity and sensitivity of automated gene function prediction via the coupling of co-expression networks and the guilt-by-association principle. Automated gene function prediction is particularly needed for the model plant *Arabidopsis* in which only half of genes have been functionally annotated based on sequence similarity to known genes. The results highlight the need for structured representation of species-specific biological knowledge, and warrants caution in the design principles employed in future ontologies.

Keywords: *Arabidopsis thaliana*, design principles of ontologies, gene function prediction, Gene Ontology, information content, MapMan

INTRODUCTION

With the ever increasing availability and quality of high-throughput data from all levels of cellular organization (e.g., transcriptome, proteome, and metabolome), ontologies have become an integral part of multivariate data analysis to facilitate biological interpretations. Accumulated knowledge in biology, unlike other scientific fields, is rather difficult to capture, and convey with mathematical formalisms. Nevertheless, ontologies offer the means for structured representation of knowledge gathered in various (electronic) written forms (e.g., text books, journal articles, databases), whereby the structure pertains to the relationships between knowledge concepts. Since ontologies are intended to represent corpora of knowledge, often in a particular field, the considered concepts can be used to annotate entities from the field of research.

Decade-long research efforts in this area, including annotation schemes such as the MIPS functional categories as well as the KEGG ontology (Ruepp et al., 2004), have resulted in ontologies tailored to different aspects of biological research, from genes and pathways to species-specific tissues, organs, and entire anatomies (Bard and Rhee, 2004). Two aspects of using biological ontologies have already been adequately addressed and thoroughly investigated, namely: (1) statistical tests for enrichment of ontological concepts (Rivals et al., 2007), (2) categorization and choice of semantic similarity measures for comparison of ontological concepts (Guzzi et al., 2011). However, the integration of biological ontologies, to facilitate interoperability of genomic databases, and their comparison, with the aim of selecting suitable ontologies, can still be regarded as pressing issues in bioinformatics and computational biology (Stein, 2003; Punta and Ofran, 2008).

In combination with methods from multivariate data analysis (e.g., clustering and separation), structured biological knowledge allows for automated reasoning and statistically sound inferences in biology. This is particularly relevant due to the recent surge of methods and applications in network-driven co-expression analysis of transcriptomics (i.e., gene expression) data. Co-expression networks provide the medium for transfer of gene annotation following the guilt-by-association (GBA) principle, whereby known (and enriched) function in a set of genes is propagated to the genes of unknown function in the set. Solutions for automated gene function annotation are still relevant even for well-investigated model organisms, such as *Arabidopsis thaliana* (*Arabidopsis*) with ∼27,000 genes of which only half have been functionally annotated based on sequence similarity to known genes, while the function

of mere 13% has been experimentally confirmed (Lamesch et al., 2012).

In modern plant biology, there are two widely used ontologies: the Gene Ontology (GO) and MapMan. While the general GO has originated as species-unspecific, MapMan was initially specifically tailored to *Arabidopsis*. Furthermore, the latter has been extended to cover other plants such as maize (Doehlemann et al., 2008), *Medicago* (Tellström et al., 2007), tomato (Urbanczyk-Wochniak et al., 2006), and potato (Rotter et al., 2007). With respect to the nomenclature of concepts, the MapMan ontology comprises a set of 34 tree-structured bins, describing the central metabolism as well as other cellular processes (e.g., stress responses). On the other hand, GO is a collection of concepts, called terms, which are connected via *is a* and *part of* relations aimed at functionally categorizing genes (for details of scope and structure of GO, the reader is directed to, Ashburner, 2000; Stevens et al., 2000; Blake and Harris, 2002). Moreover, GO can be regarded as a collection of three ontologies that correspond to independent categories of gene function: molecular function (GO-MF), biological processes (GO-BP), and cellular component (GO-CC). Functional categorization of genes can also be performed across species with the help of high-level GO terms, reducing GO to the so-called GO slim ontology. Besides the generic species-unspecific version, there are GO slim ontologies which are designed for specific species, e.g., *Saccharomyces cerevisiae* (Cherry et al., 2012), *Arabidopsis* (Lamesch et al., 2012), and *Drosophila* (Adams et al., 2000). In MapMan, the original assignment of bins was based on publicly available gene annotation in TIGR (The Institute for Genomic Research), adopting a process alternating between automatic recruitment, and manual correction (Thimm et al., 2004).

Although the two ontologies have both been used in plant research, systematic comparison of GO and MapMan has not yet been undertaken. Assessing the advantages and drawbacks of the two is crucial for the selection of the ontology suitable for automated gene function annotation. Here we present the findings from the comparative analysis of GO and MapMan, first by analyzing similarities and differences with respect to the (1) overall structure and size, and (2) design principles. Here, we suggest suitable preprocessing strategies to alleviate the problem of inconsistent mappings regarding the inheritance of concepts given by the respective structure of the ontology.

Furthermore, for the specific case of the gene annotation for *Arabidopsis*, we investigate (3) the coverage and (4) biological relevance of concepts within the two ontologies. In addition, we analyze the effect of a particular ontology on the function transfer across genes based on the coupling between the GBA principle and co-expression networks. The findings from our comparative analysis point out that the domain in which ontologies are used may have a profound effect on the selection of a best-performing alternative. Therefore, our results pinpoint the need for development of methods for objective, systematic, and problem-specific comparison of biological ontologies as well as formal frameworks for transfer of ontologies in cross-species analyses.

RESULTS

THE STRUCTURE OF MAPMAN AND GO

Although the relationships between two ontological terms in GO and MapMan can be described by *is a* and *part of* relationships, the structures of the two ontologies differ. While all three categories of GO are structured in the form of a directed acyclic graph (DAG; Yon Rhee et al., 2008), the relationships in MapMan are modeled following a tree structure (cf. **Figure 1**). The implication of using a DAG as an underlying structure of the ontology is that child concepts may have more than one parent. The multiplicity of parent concepts can be regarded as an advantage, as it provides a high degree of flexibility and may enable powerful grouping, searching, and analysis of genes (Yon Rhee et al., 2008). In contrast, although the tree structure closely resembles the intuitive connotation of a hierarchy of concepts, it sacrifices a part of the flexibility when the ontology is updated (e.g., by adding new concepts).

A disadvantage of the DAG structure, compared to a tree, is that the depth of a concept cannot be unambiguously defined, since there may exist multiple paths to the root node. Therefore, we define the depth of a concept in GO (i.e., term) as the shortest path to the root node, corresponding to the minimum concept depth (see Guzzi et al., 2011) for other similar measures). In addition, multiple parent concepts increase the overall number of possible ancestors at the same concept depth. This is particularly the case when comparing the DAG structure of GO with the tree structure of MapMan. Furthermore, the number of potential parent concepts as well as the overall size of an ontology renders it difficult to visualize concept associations for large-scale transcriptomic analyses (for the plethora of available visualization methods see, e.g., Zeeberg et al., 2003; Tsiaras et al., 2008; Carbon et al., 2009; and has effect on statistical hypothesis testing, e.g., in multiple testing scenarios Goeman and Mansmann, 2008).

An immediate solution represents GO slim, which categorizes genes on the basis of a relatively small set of high-level GO terms. Like in the tree structure of MapMan, the smaller number of (parent-) terms of the slim ontologies facilitates the interpretability of obtained results. However, similarly to the previous arguments, a small number of parent terms can also turn out to be a disadvantage, as it may lead to a comparatively flatter hierarchy structure, regardless of the actual size of the used ontology. Subsequently, a flat hierarchy may compromise the specificity and biological relevance of individual concepts due to its coarseness.

DESIGN PRINCIPLES OF ONTOLOGIES – CAPTURING BIOLOGICAL CONCEPTS

An important characteristic of GO is the division in three non-overlapping domains of molecular biology–biological process (GO-BP), molecular function (GO-MF), and cellular component (GO-CC; Ashburner, 2000; Harris and Gene Ontology, 2004). While terms in GO-BP domain describe biological objectives and processes in which the annotated genes participate, terms in GO-MF characterize biochemical activities that ultimately contribute to biological processes. Finally, GO-CC summarizes the subcellular localization where a gene product is active. In contrast, while MapMan does not have a structure composed of independent categories, one can still distinguish between high- and low-level bins. Since the design principle of MapMan was to intuitively characterize and visualize metabolic pathways and processes (Thimm et al., 2004), high-level bins tend to be similar to terms in the GO-BP ontology, whereas low-level bins often resemble terms from the GO-MF ontology.

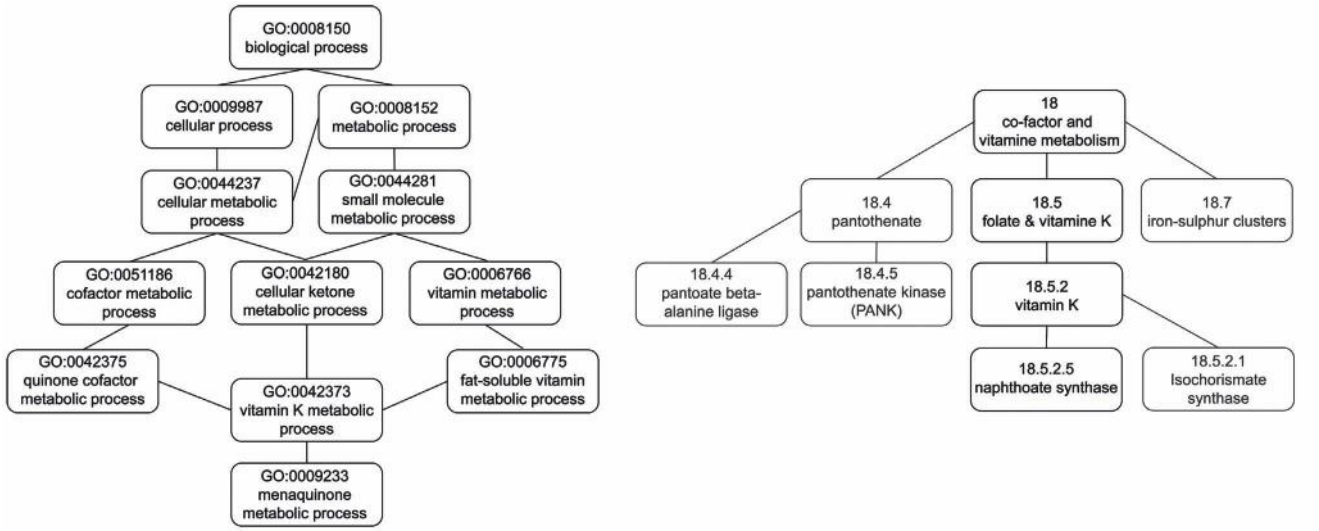


FIGURE 1 | Comparison of the concepts used in the annotation of the gene coding for naphthoate synthase (at1g60550, 264920_at) in the DAG and tree structure of GO DAG and MapMan, respectively.

To illustrate this claim based on the whole annotation of gene products rather than examples of individual concepts, we quantified the similarity of MapMan bins and GO terms by utilizing a network-based approach. For the purpose of this analysis, nodes correspond to concepts, i.e., terms in GO and bins in MapMan. An edge between two nodes is established if the set of genes which are annotated with the respective terms corresponding to the nodes are similar (cf. Materials and Methods).

Figure 2 shows the resulting network which consists of all GO-MF and GO-BP terms that exhibit a similarity to at least one MapMan bin. The edges of the resulting network can further be divided by the type of association they model, namely: similarity between MapMan bin and GO-BP term, MapMan bin and GO-MF term as well MapMan bin, and both GO-MF and GO-BP terms. Inspection of the three types of edges in this concept-association network shows that high-level MapMan bins are often associated with terms originating from GO-BP. In contrast, MapMan bins deeper in the hierarchy are predominantly associated with GO-MF terms. A statistical analysis quantifies this observation as the difference of average depth of concepts for the first two of the groups of edges is statistically significant at the 5% level (Wilcoxon-Rank-Sum test, p-value $= 0.016$, cf. **Figure 3**). Here and in the following, we only use the terms from the two GO ontologies, namely: GO-MF and GO-BP, since there is no correspondence between GO-CC and any bin in MapMan.

GENE ANNOTATION COVERAGE – THE STATUS QUO FOR *ARABIDOPSIS*

The genome of *Arabidopsis* contains 27,416 protein coding genes according to the latest genome annotation version (TAIR10, November 2010)[1] which excludes pseudo genes and genes encoded by transposable elements (Lamesch et al., 2012). Inspection of these mappings shows that a total of 15,238 gene products are annotated with MapMan bins, while 12,225 and 13,157 genes are annotated by GO-BP and GO-MF terms, respectively. By combining the available annotation of all three ontologies ∼63% of *Arabidopsis*' genes can be annotated.

The number of genes that are annotated with both MapMan and GO terms (either GO-BP or GO-MF) is ∼87% of the total number of annotated genes with concepts from any of the three ontologies (cf. **Figure 4**). Furthermore, each ontology contains concepts used in the annotation of a unique set of genes: the contribution of MapMan is slightly larger, with 2,557 unique bins, compared to 625 and 572 terms for GO-MF and GO-BP, respectively (**Figure 4**). In summary, the coverage of the two ontologies is comparable, which further serves as a justification for the undertaken comparative analysis.

In addition, we find that 3,598 unique GO-BP terms are used to annotate ∼45% of *Arabidopsis*' genes. GO-MF contains 2,148 unique terms covering ∼48% of the genes. Finally, 1,361 unique bins of MapMan are used in annotating 56% of *Arabidopsis*' genes. Similarly to the overall size of the ontologies, we demonstrate that the average number of parent terms per gene in MapMan is 3 in comparison to 20 and 7 in GO-BP and GO-MF, respectively. Clearly, MapMan is the smaller ontology with roughly one-third of the size of GO-BP.

Furthermore, to see whether a comparatively low number of parent terms ultimately results in an overall flatter hierarchy structure in the case of MapMan, we analyze the differences in the distribution of depth in the two ontologies. Again, we contrasted the concept depth distribution on the current state of ontological gene annotation in *Arabidopsis*. Here, for each annotated gene, we determined the depth of every associated term and all of its parents (further defined as "complete ontology", see Materials and Methods). As shown in **Figure 5**, MapMan indeed represents a flatter hierarchy: while both term depth distributions of the two GO categories closely resemble a normal distribution with a mean $\cong$ median term depth of ∼5 (sample skewness:

[1] http://arabidopsis.org

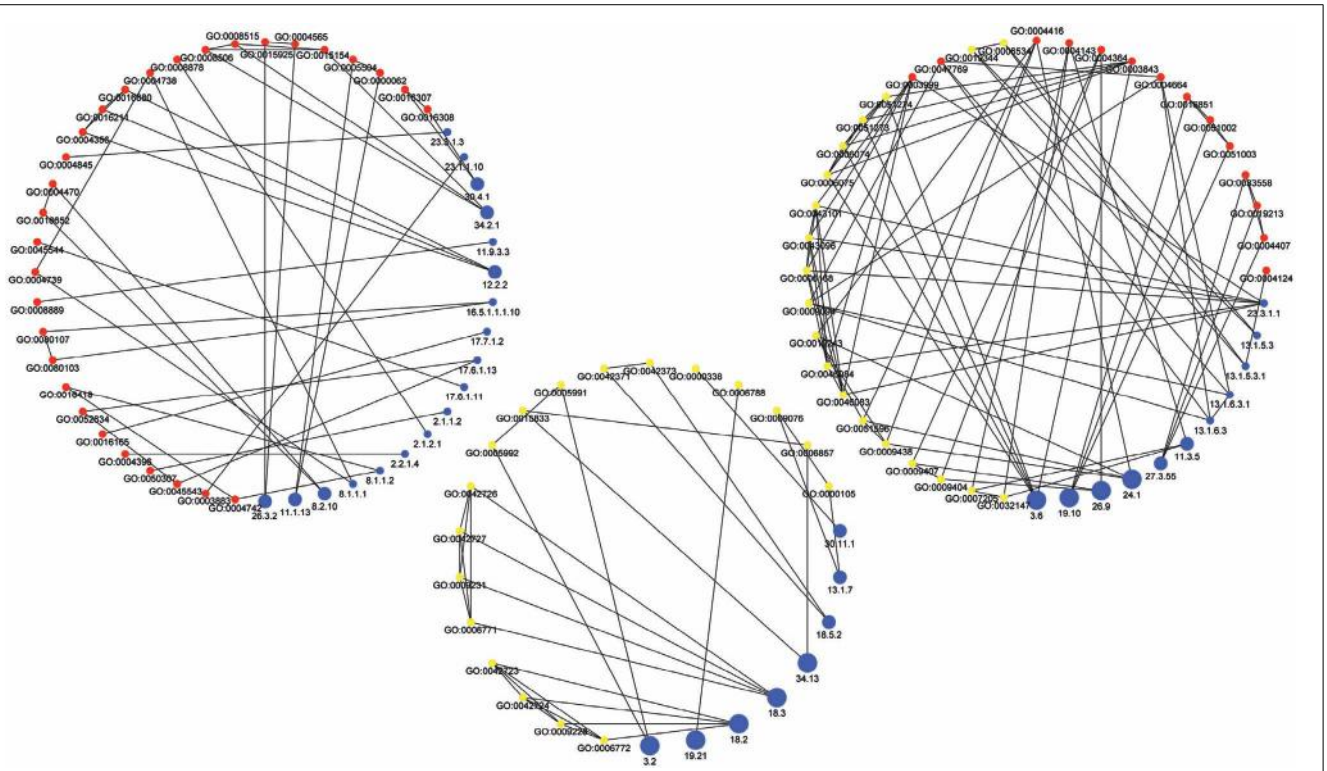


FIGURE 2 | Network of associations between MapMan bins and GO terms. Blue nodes represent MapMan bins, yellow denote GO-BP terms, and orange nodes correspond to GO-MF terms. The size of the nodes representing MapMan bins corresponds to their depth in the ontology. Note that the low-level bins (small blue nodes) correspond to terms in GO-MF, while high-level bins (large blue nodes) are associated to terms in GO-BP. The edges of the network are divided by the type of association they model, namely: association between MapMan bin and GO-BP term (left), MapMan bin and GO-MF term (middle) as well MapMan bin and both GO-MF and GO-BP terms (right).

GO-MF = 0.01,GO-BP = 0.26), the term depth distribution of MapMan is skewed toward lower values with median term depth of three (sample skewness: 0.69). In addition, the maximum term depth is lower, and is of value seven in MapMan and 10 in both GO categories, respectively.

INFORMATION CONTENT OF ONTOLOGICAL TERMS

Common to both ontologies is that high-level concepts describe general processes, functions, or structures, while low-level concepts are more specific. The previous claim that MapMan constitutes a flatter hierarchy structure, compared to GO, needs further investigation to ascertain whether the structure of MapMan can be used equally well in elucidating biologically meaningful information from its ontological concepts (i.e., bins). Here we rely on the information content (IC) of an ontology concept to quantify its specificity by accounting for the overall number of genes annotated with it (Resnik, 1995). Briefly, the information content of an ontology concept is lower as its specificity decreases; the more abstract a concept, or broader an ontological category, the lower its information content (see Material and Methods). **Figure 6** shows a histogram of the IC of all MapMan, GO-MF and GO-BP used in the annotation of the *Arabidopsis*' genome. One can observe that both GO ontologies exhibit a higher maximum IC as well as more terms of large IC. Moreover, the median IC of 9.23 for MapMan is smaller than that of GO ontologies, i.e., 10.4 for GO-MF and 10.82 for GO-BP. This implies a slightly coarser grouping of processes and functions in the case of MapMan. However, one can also observe that MapMan contains more terms of average IC (∼5.5 to ∼8.5).

Besides the analysis of the distribution of ICs for concepts of an ontology, it is important to also investigate the interplay between the underlying structure (captured by the concept depth) and IC to characterize the level at which a deeper hierarchy relates to more specific sets of genes. This dependence between concept depth and IC is visualized in **Figure 7** with the help of box plots. One can observe that all three ontologies exhibit an asymptotic trend of the median IC values per concept depth. Interestingly, none of the ontologies displays a gradual trend of a linearly increasing IC with the increasing concept depth. Further, this non-linear behavior can be modeled using classical Michaelis-Menten kinetics (Lehninger et al., 2008), which relates the rate of a reaction (dependent variable) with the (saturating) concentration of its substrate (independent variable). The relation is fully described by two parameters: $V_{\max}$, representing the maximum rate achieved at maximum (saturating) substrate concentrations, and K_{m}, denoting the substrate concentration at which the rate is half of $V_{\max}$. Analogously to this classical enzyme kinetics, we take $V_{\max}$ to denote maximum IC achieved at maximum concept depth and K_{m}, the concept depth at which the IC is $V_{\max}/2$. By using non-linear (least-squares) regression (Leskovac, 2003), we

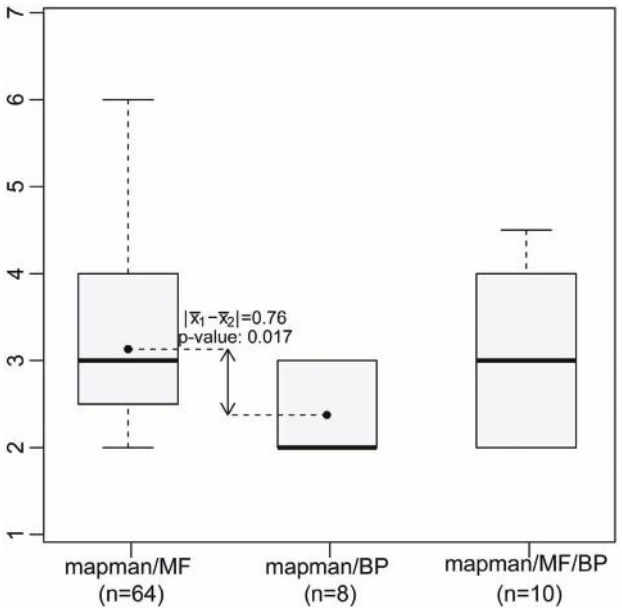


FIGURE 3 | Concept depth of MapMan bins associated with GO terms based on similarity of the annotated genes (cf. Figure 2). Low-level MapMan bins tend to represent terms from the GO-MF ontology (left boxplot), while high-level MapMan bins are associated with terms from GO-BP (middle). The difference of the bin depth for these two groups is statistically significant at level $\alpha = 0.05$.

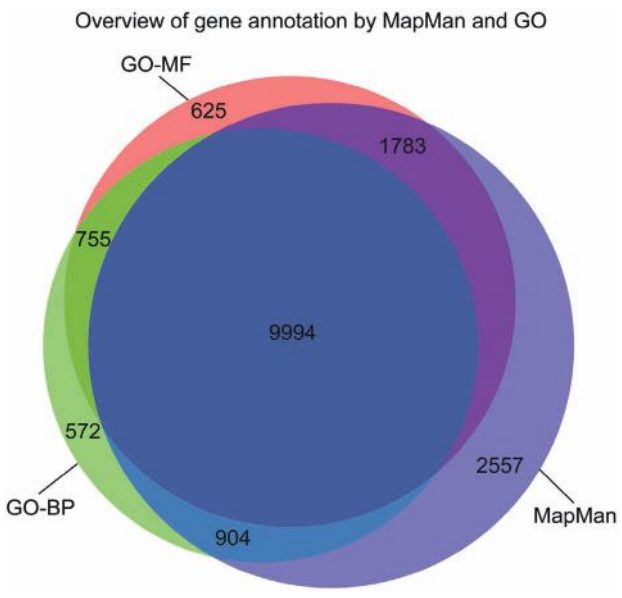


FIGURE 4 | Venn diagram illustrating the overlap in the gene annotation of Arabidopsis thaliana with concepts from MapMan, GO-BP, and GO-MF, respectively.

obtain estimates for the constants V_{max} and K_m (cf. Materials and Methods). Interestingly, we find the all determined K_m values are close to ∼1, relating to $V_{max}/2$ of 6.08, 6.04 and 6.76 for MapMan, GO-MF and GO-BP, respectively. Therefore, we conclude that, in the case of *Arabidopsis*, all three ontologies possess the similar structural capabilities to allow for an adequate biologically meaningful discrimination of concepts and genes.

EMPLOYING MAPMAN AND GO FOR AUTOMATED GENE FUNCTION ANNOTATION – THE CASE STUDY OF *ARABIDOPSIS*

The current incompleteness of available gene annotation for *Arabidopsis* clearly emphasizes the need for automated gene function prediction, even in the case of well-studied model organism. In addition to sequence similarity, gene co-expression analysis employing genome-wide transcriptomics data across tissues or in response to environmental perturbation has become a valuable tool to predict gene function based on the GBA principle (Klie et al., 2010). The transfer of function annotations between two genes, exhibiting similar profiles, according to GBA is now a standard procedure for gene function prediction. Furthermore, gene co-expression networks have emerged as a powerful representative of the structure of similarity of transcriptomic profiles, and are readily employed for intra-species transfer of gene annotations following GBA (e.g., in the field of plant science see, Obayashi et al., 2009; Mutwil et al., 2010; Mochida et al., 2011).

Due to the previously described difference in the structure of GO and MapMan, we next evaluate the effect of these characteristics on the performance of gene function prediction by using a GBA-based network-driven approach. To this end, we employ a transcriptomic data-set of 273 publicly available *Arabidopsis* microarray experiments to construct a gene co-expression network (see Materials and Methods). We rely on the approach described in Mutwil et al. (2011) to obtain a co-expression network which is based on robust statistical parameter estimation combined with successive optimization of the biological relevance of the obtained network. In the co-expression network, the nodes correspond to *Arabidopsis*' genes, and edges are established if the incident nodes (i.e., genes) are mutually in the top 30 most similar genes. The similarity is assessed by the Pearson correlation coefficient, and this approach, termed highest reciprocal rank, has already been characterized to optimally capture functional annotation of co-expressed genes (Obayashi and Kinoshita, 2009).

In the following, we rigorously extend this method to allow for a network-based prediction method of gene annotation by employing the method of majority voting (cf. Materials and Methods). In majority voting, the annotations of all adjacent nodes (i.e., immediate neighbors) of a given gene are ordered in a list, from the most to the least frequently appearing (Schiwikowski et al., 2000). The function of an unannotated gene is then predicted by the first k functions in the list. Note that k is a user-specified parameter. Although the approach is very simple, it is exceptionally fast and can serve as an excellent reference for the amount of local information captured by the network due to the consideration of annotations of immediate neighbors.

To generate and verify predictions of annotation with both ontologies, we conduct the following simulation: We first select the genes which are annotated with concepts from each of the three ontologies, i.e., MapMan, GO-MF, and GO-BP, which resulted in 9,994 genes. Moreover, the annotation provided in all three ontologies for a set of randomly chosen genes is discarded. To

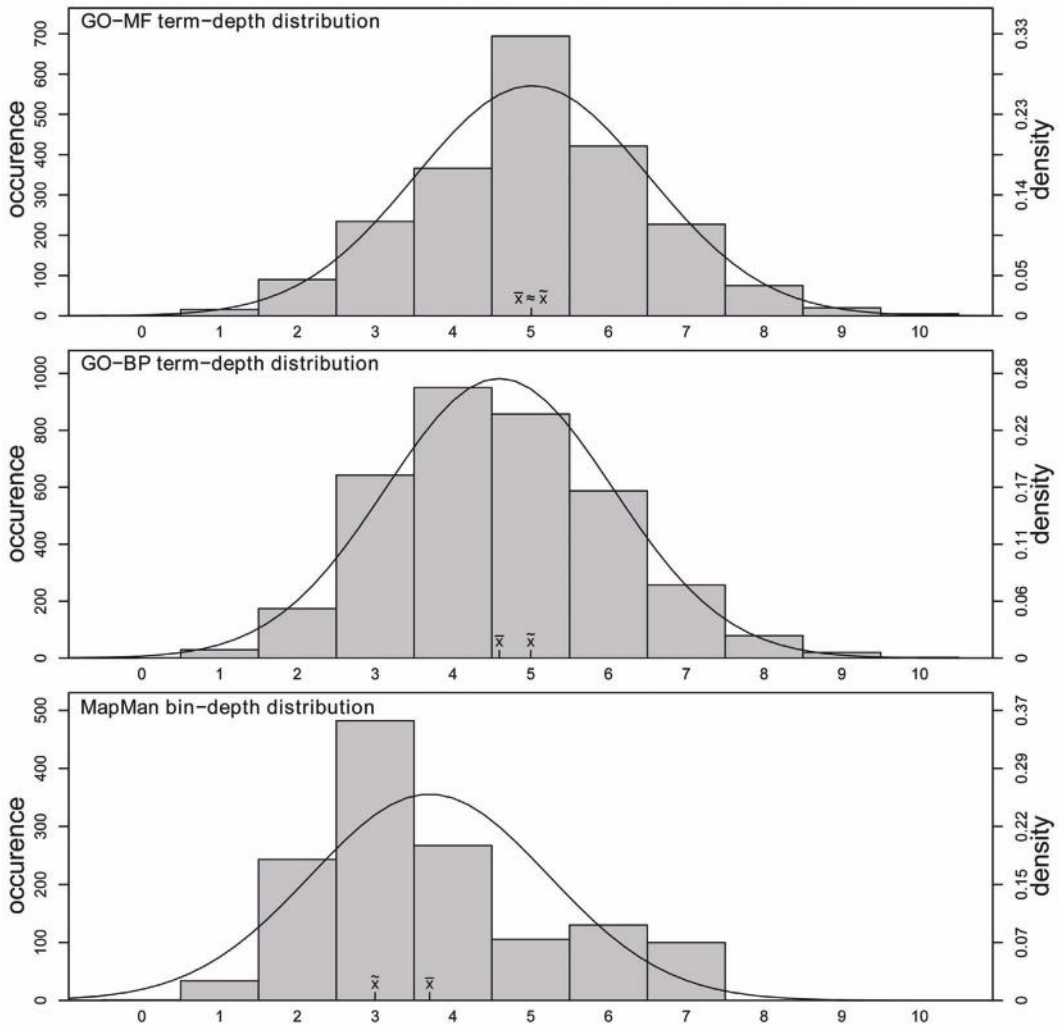


FIGURE 5 | Distributions of concept depth in GO-MF (upper), GO-BP (middle), and MapMan (lower). The *x*-axis denotes the depth of a concept while the left *y*-axis denotes the corresponding occurrence. For all three distributions, a normal distribution is fitted (right *y*-axis).

this end, the number of this artificially unannotated genes is set to be 4,000, corresponding to a fraction of ∼40% genes of unknown function. This scenario closely resembles the current state of *Arabidopsis*' gene annotation. For these genes, prediction of gene annotation is obtained by using each one of the three ontologies. The predictions of the top $k \in [1,20]$ most abundant concepts in the network vicinity are evaluated for their performance based on the original discarded annotation. For every k most abundant concepts from the unannotated genes, this procedure is repeated 1,000 times, such that in every iteration a different set of randomly unannotated genes is sampled. Note, that all three used ontologies were preprocessed so that for each gene all parent terms are included. Moreover, to avoid trivially correct predictions, such as the root terms of GO-MF and GO-BP, we do not consider the root terms as well the 20 less informative terms (based on the IC; see Materials and Methods). The predictions are summarized by precision and recall, two widely used performance measures in information retrieval and binary classification (Baeza-Yates and Ribeiro-Neto, 1999), as well as by their harmonic mean, the F-measure. On the other hand, we evaluate the biological relevance of the obtained predictions by investigating the normalized IC (with respect to the maximum) and the depth of the top k, $k \in [1,20]$ predicted terms. Additionally, we also report the number of genes for which a prediction can be obtained following this procedure.

Figure 8 summarizes the acquired prediction performance results for all three employed ontologies. One can observe that the use of MapMan exhibits an advantage in the performance of gene function prediction, as the combined F-measure is the highest over the whole range of top k, $k \in [1,20]$ concepts (the exception is the case of $k = 20$, where the F-measure is zero, due to the lower number of terms in MapMan). This is mainly due to a higher average recall, i.e., a higher fraction of all the originally concepts, used in the annotation of a gene, that were successfully retrieved. Nevertheless, the average precision between GO and MapMan is comparable, indicating that the ratio of correctly

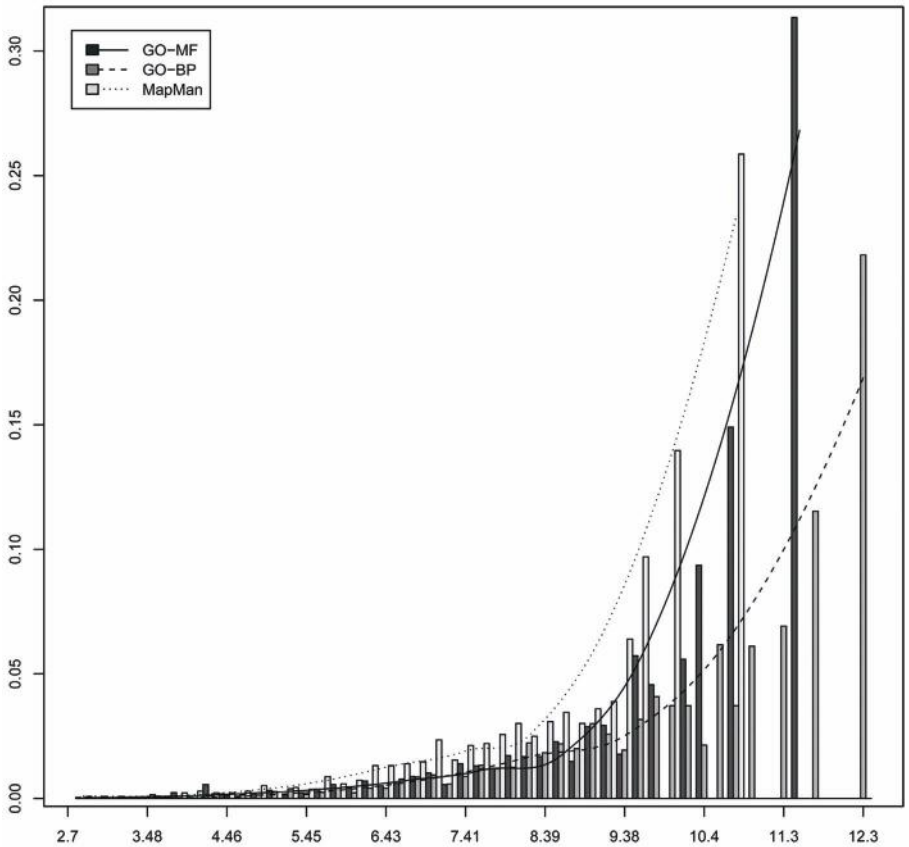


FIGURE 6 | Histogram of the information content of all concepts used to annotate Arabidopsis' genome by using the three ontologies MapMan, GO-MF, and GO-BP.

predicted concepts to all predicted concepts is similar across all three ontologies.

Correspondingly, the average IC and depth of concepts is generally higher in the case of MapMan, which implies a higher biological relevance or specificity of the predicted terms (**Figure 8**). However, both GO ontologies perform better with respect to the fraction of genes for which any prediction of gene annotation can be derived, i.e., 51% for MapMan vs. 64 and 73% for GO-MF and GO-BP, respectively. This suggests that the distribution of genome annotation is less clustered and more homogeneous.

DISCUSSION

Here, we provided the first comparative analysis of two ontologies, GO, and MapMan, both widely used in plant biology studies. The first part of the comparison comprises the structural characteristics of the ontologies, namely: the type of concepts and relationships between them as well as the design principles underlying GO and MapMan. Our findings were in support of the claim that higher level bins in MapMan correspond to terms of GO-BP, while lower level bins are more similar to terms of GO-MF. Regardless of these analogies, GO offers the possibility to also investigate gene products with respect to their spatial distributions, captured in the terms of the third GO ontology – cellular component (GO-CC). In contrast, MapMan does not facilitate spatial analysis of genes and the downstream processes (e.g., metabolism). Nevertheless, although cellular processes and molecular functions are represented well in both GO and MapMan, temporal changes during plant development, fruit ripening, or progression of stress are in their nascent stages. Therefore, future developments in plant-specific ontologies should consider integrating the indicated spatial and temporal dimensions indispensable for accurate description of molecular processes in plants.

In the second part of the study, we investigated the annotation corpus of *Arabidopsis*' genes and carried out a detailed comparison of the two ontologies with respect to the information content of the respective concepts, i.e., bins in MapMan and terms in GO. It turned out that MapMan, GO-BP, and GO-MF exhibited similar relationships between information content and depth of concepts. In conjunction with the plethora of existing tools for computational analyses based on both ontologies, our results indicated that both ontologies may be equally suitable with respect to the biologically meaningful information that could potentially be extracted.

Finally, we used the two ontologies as a principle source of information in the context of automated gene function prediction following the GBA principle on co-expression networks. The co-expression networks were created by using publicly

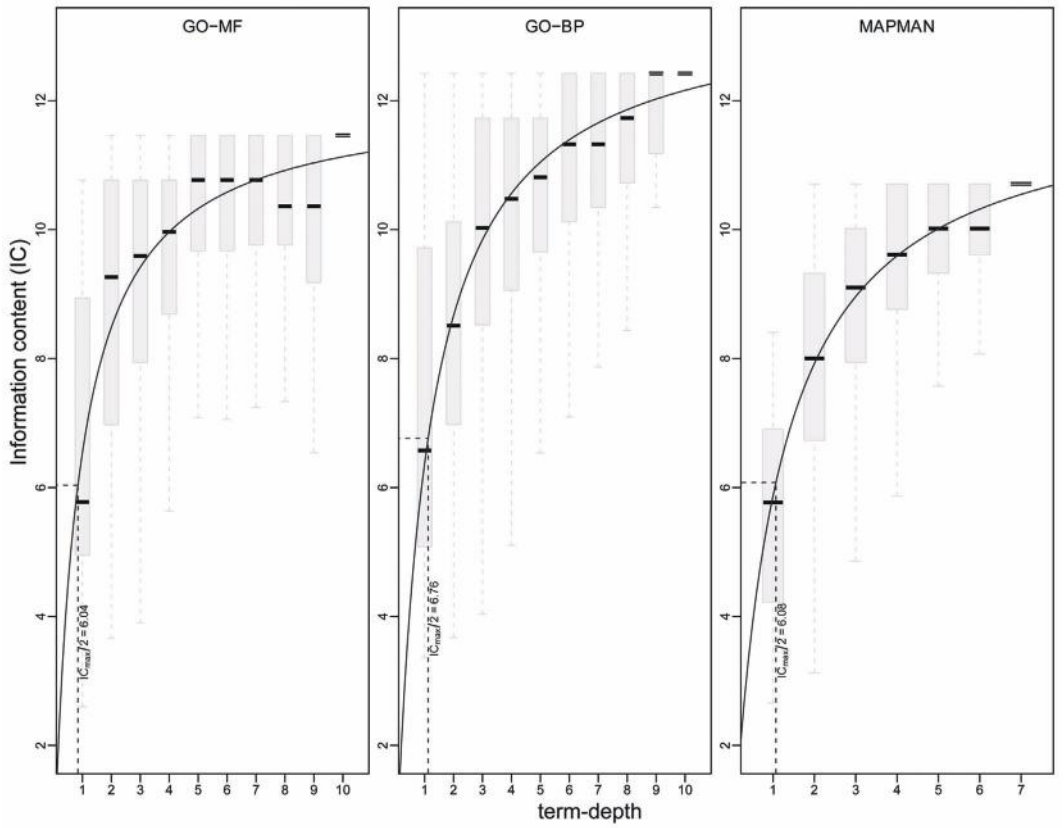


FIGURE 7 | Visualization of information content (*y*-axis) at a given term depth (*x*-axis) for GO-MF (left), GO-BP (middle), and MapMan (right). A non-linear regression is used on the medians of the information content per concept depth following Michaelis-Menten kinetics (solid line). The constant $V_{max}/2$ is shown by dotted lines (see main text for details).

available transcriptomics data sets for *Arabidopsis*, and provided the medium for local propagation of concepts to unannotated genes in the vicinity of a given well-characterized gene. To this end, we used the simplest available alternative for automated function annotation given by the majority voting. Although our findings that MapMan outperformed GO with respect to function annotation depend on the algorithm for annotation transfer, we believe that they are robust as most of the available algorithms rely on propagation of local information only. While MapMan's tree hierarchy at a first glance appears to be flatter, as assessed by term depth, and IC, in comparison to GO's DAG structure, MapMan's design tailored to *Arabidopsis* is most likely reflected in the improved performance in gene function prediction. In contrast to MapMan, GO represents a more generic ontology, reflected in its changing structure and gene annotation. Since no other plant model organism is currently equally well-annotated by GO and MapMan as it is the case for *Arabidopsis*, no general conclusions for plant species can be made. Nevertheless, what remains to be investigated is the effect of the distribution of annotated genes in the network. In other words, we expect that choice of the ontology for automated gene function annotation will ultimately depend on the dispersion of patches of annotated nodes (following the focused biological interest in genes of particular process/function).

Last but not the least, the major implication of our study is that the choice of which ontology to be used computational analyses is problem-specific, as it highly depends on the interplay between the structural properties of the ontology, the size, and quality of the annotation corpus, using the ontology, as well as the employed multivariate data. Therefore, we believe that aside from the comparison of ontologies based on intra-ontology characteristics (e.g., distribution of information content), our study emphasizes the need for another criterion – namely, the biological question to be answered by using ontologies, for instance, comparison of plant developmental stages, or plant-specific structures and the here addressed gene annotation. This, of course, may open yet another field of bioinformatics research related to the design of sound methods for ontology selection suitable for a particular problem at hand. In this respect, we believe that the suggested direction may result in development of (external and internal) measures for problem-specific comparison of ontologies and their performance – an issue which was already addressed in other research areas (e.g., data clustering, retrieval in audio and video databases). Taken altogether, the identified issues warrant caution in extending ontologies from model to other species and suggest that this may be most appropriately performed in a careful semi-automated manner.

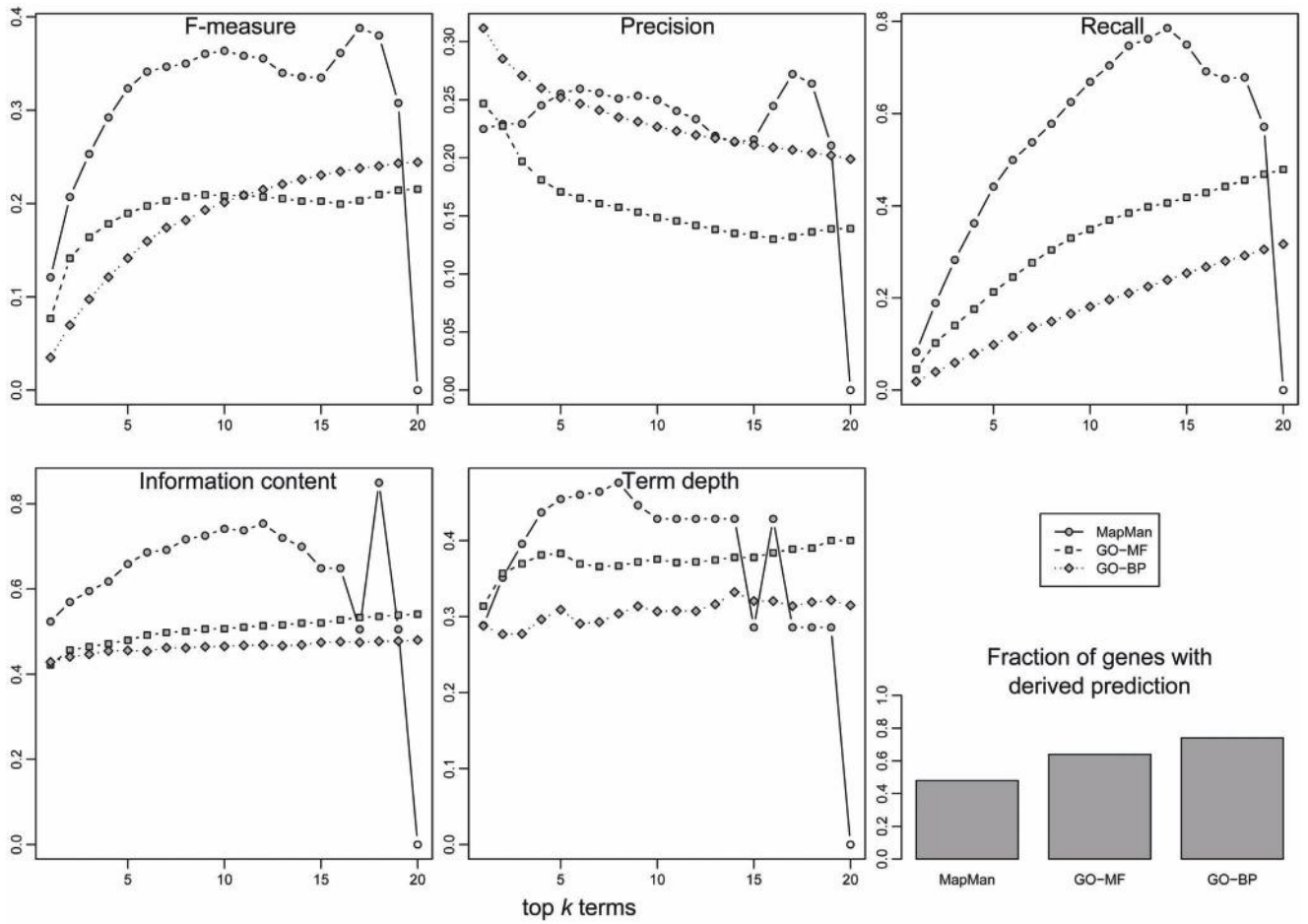


FIGURE 8 | Evaluation of prediction performance (*y*-axis) for gene function annotation using the three ontologies. Three classical performance indices from information retrieval were employed: F-measure (upper-left), precision (upper-middle), and recall (upper-right). In addition, the information content (lower-right), and the concept depth relative to the maximum depth within the respective ontology (lower-middle) of predicted concepts were evaluated. In all cases, the 20 most abundant concepts (*x*-axis) in an unknown gene's neighborhood are evaluated. Finally, the fraction of unknown genes for which a prediction could be derived is given (bar plot, lower-right).

MATERIALS AND METHODS

ARABIDOPSIS TRANSCRIPTOMICS DATA-SET AND RECONSTRUCTION OF A GENE CO-EXPRESSION NETWORK

The employed transcriptomic data-set used to derive the gene co-expression network consist of 279 of publicly available microarray experiments (Affymetrix Ath1 gene-chip, 22,500 probe sets) obtained from the Gene Expression Omnibus[2] (Edgar et al., 2002). Note, that this is the same transcriptomics compendium which is used in the PlaNet co-expression analysis platform (Mutwil et al., 2011). Initially, a total of over 6,000 microarray experiments were downloaded and the quality of each individual microarray experiment was ensured by an automated outlier detection and quality control. Here, the R Bioconductor package `arrayQualityMetrics` (Kauffmann et al., 2009) was employed to conduct (1) between-array comparisons based on distance between arrays and Principal Component Analysis, (2) inspection of array-wide probe intensity distributions by boxplots and density plots, (3) variance-mean dependence of each array, and (4) individual array quality assessment by MA plots. After this preprocessing, 1,707 microarrays were retained. Furthermore, this transcriptomics compendium was reduced by selecting a subset of experiments comprising 273 microarrays. This is performed to remove any bias arising through (potentially) un-informative or repetitive data while preserving the overall structure of the transcriptomics compendium (Mutwil et al., 2011). Briefly, this selection strategy is based on the Subset Selection problem from linear algebra, whereby, for a given number l and a matrix A, one is to find the subset of l columns from A which are most mutually independent. Here, columns of the matrix A denote individual microarray experiments (1,707 in total), rows correspond to genes, such that each matrix entry represents the corresponding gene expression levels. Application of the outlined selection procedure yielded 279 microarrays which were subsequently normalized using quantile normalization via the `simpleAffy` R package. This data-set was used to reconstruct

[2]http://www.ncbi.nlm.nih.gov/geo/

the co-expression network and is available in the Table S1 in Supplementary Material.

PREPROCESSING OF ONTOLOGIES – REMOVAL OF INCONSISTENCIES AND INTEGRATION OF PARENT CONCEPTS

As sources of mapping genes to ontology terms in *Arabidopsis*, we employed the latest versions available for MapMan (Version 1.1 from January 2010)[3] and GO (Version 2.5 from September 2010, available via the R package `ath1121501.db`[4]). Within these mappings, a total of 15,238 gene products are annotated with MapMan bins and 12,225 and 13,157 genes are annotated by GO-BP and GO-MF terms, respectively. However those raw mapping files contain inconsistencies: while the annotations for some genes contain only the most specific concepts, i.e., terminal or leaf concepts with no further child concepts, others are additionally annotated with parent concepts. As an example, consider the genes annotated with the MapMan bin "29.5.11.4.2" corresponding to "protein.degradation.ubiquitin.E3.RING" in *Arabidopsis*. This bin is a leaf or terminal concept, i.e., it has no children. One gene that is annotated with this concept is a member of the ARM repeat superfamily (locus ID at1g71020) and is additionally annotated with the parent bin "29.5.11" corresponding to "protein degradation ubiquitin." However, other genes annotated with the bin 29.5.11.4.2, for instance *EDA40* (at4g37890), are only annotated with the leaf bin "29.5.11.4.2" missing the mapping to any parent bins, e.g., 29.5.11.4 or 29.5.11. Likewise, similar examples hold for both GO domains, GO-MF, and GO-BP. In total, 25 of such inconsistencies can be identified for MapMan and 3,750 and 2,202 for GO-MF and GO-BP, respectively.

The effect of an incomplete mapping which includes only partially – or even not at all – parent concepts is twofold: first, the analysis by means of IC of a concept would lead to incorrect results since the IC of a concept is dependent on the number of genes associated with it. By definition of an ontology, a gene annotated with a low-level concept should automatically be annotated with all of the ancestral terms, too (**Figures 1** and **9**). Only considering the concept-gene association counts in a raw ontology will lead accidentally to erroneous results for the derived ICs; in this case leaf or terminal concepts might exhibit a higher IC than their parent terms (Klie et al., 2010). Second, for the purpose of gene function prediction in majority voting, common ancestor terms of the neighboring genes are of great importance. In the case that the annotation of all neighboring genes is a disjoint set of low-level concepts, no majority vote can be found (cf. **Figure 9D**). However, the gene's neighbors can share common parent concepts that can help in deriving predictions for the gene in question. Although the derived annotation might not be as specific, the prediction of a high-level concept suggesting the putative involvement in processes or pathways is preferred to obtaining no prediction at all. To resolve the problem of incomplete mappings, we preprocessed all three ontologies so that for each gene, the complete list of parent terms is included. Note, that those parent terms can readily be identified by enumerating the respective DAG or tree structure defined by *is a* or *part of* relations (**Figures 9A,B**). We further define these modified mappings as "complete ontologies".

Finally, the preprocessing involved removal of control and unknown probe sets, which corresponds to probes associated with MapMan bins 0 and 35 ("control" and "unknown"/"not assigned") and all their child bins.

EVALUATION OF ONTOLOGY STRUCTURE AND INFORMATION CONTENT

We employ two measures to characterize the structure and the characteristics of an ontology – the depth and the information content (IC) of concepts.

Given a directed acyclic graph $G = (V, E)$, which defines the relationships of concepts within an ontology, where V is a set of vertices, E is a set of edges, the depth of a term x is given by the *distance* $d(x, r)$ between the two vertices x and r, where node r corresponds to the root concept of the ontology. Furthermore, the distance is defined as the length of the shortest path from x to r (Bondy and Murty, 2008). Note that node r represents the root term which is explicitly defined for GO-BP as and GO-MF and which can be implicitly defined for MapMan by adding an artificial root node, i.e., bin *r*.

The IC of an ontological concept c is defined as $IC(c) = -\log_2(|G_c|/|G_{\text{all}}|)$, where G_c is the set of genes annotated with the concept c and G_{all} is the set of genes annotated with any of the concepts in the ontology (Resnik, 1995).

DETERMINING SIMILAR CONCEPTS ACROSS ONTOLOGIES

To quantify the similarity of two concepts c_1 and c_2, we use the Jaccard similarity coefficient of the set of genes G_1 annotated with concept c_1 in MapMan and the set of genes G_1 annotated with concept c_2 in GO. The Jaccard similarity coefficient for two sets G_1 and G_2 is defined as $\text{sim}(c_1, c_2) = J(G_1, G_2) = |G_1 \cap G_2|/|G_1 \cup G_2|$. As 50% of all MapMan and GO concepts describe four or more genes, we consider only concepts of MapMan and GO that are annotated with at least four genes (i.e., $|G_1|$ and $|G_2| > 3$) to avoid identifying similar concepts based on individual genes.

In addition, to analyze the pair-wise similarity over all concepts, we create a network in which nodes correspond to concepts and edge are established between two nodes c_1 and c_2 if sim $(c_1, c_2) \geq 0.6$. Note, that despite its numerical value, this threshold is rather strict as it refers to only the highest 1% of all observed pair-wise concept similarities, not only between MapMan and GO but also within the respective ontologies. Nodes corresponding MapMan bins that are not connected to a node denoting a GO term are discarded. Finally, the edges of the resulting network can be divided by the type of association between nodes they model: the similarity between a MapMan bin and GO-BP term, a MapMan bin and GO-MF term as well as a MapMan bin and both GO-MF and GO-BP terms. For each of those three derived types of associations, the average bin depth of MapMan bins is determined and the statistical significance of the difference of means within the first two groups (MapMan/GO-MF, MapMan/GO-BP) is derived via Wilcoxon-Rank-Sum test (Sokal and Rohlf, 2003).

GENE FUNCTION PREDICTION USING NETWORK-BASED MAJORITY VOTING

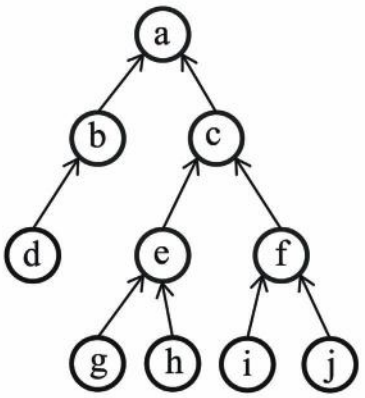
Majority voting is one of the simplest, yet fastest, network-based gene function prediction methods (Schiwikowski et al., 2000).

[3] http://mapman.gabipd.org/
[4] http://www.bioconductor.org

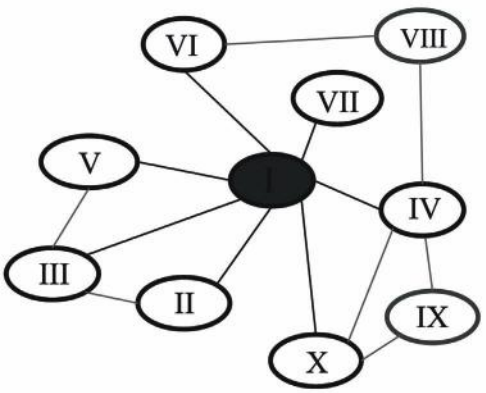
A

gene	"raw" annotation	"completed" annotation
I	*unknown*	*unknown*
II	{g}	{a,c,**e**,**g**}
III	{h}	{a,c,**e**,**h**}
IV	{i}	{a,c,**f**,**i**}
V	{f,c}	{a,c,**f**}
VI	{j}	{a,c,**f**,**j**}
VII	{e}	{a,c,**e**}
VIII	{b,e,h}	{a,b,c,**e**}
IX	{a,c,f}	{a,c,**f**}
X	{d}	{a,b,**d**}

B



C



D

term	"raw" annotation counts	"completed" annotation counts	"completed" annotation counts *filtered*
a	0	**7**	0
b	0	1	0
c	**1**	**6**	0
d	**1**	1	1
e	**1**	3	**3**
f	**1**	3	**3**
g	**1**	1	1
h	**1**	1	1
i	**1**	1	1
j	**1**	1	1

FIGURE 9 | Preprocessing of ontologies and network-based gene function prediction by majority voting. (A) The original annotation of genes ("raw" annotation) and corresponding concepts (denoted by letters) is extended for each gene by including all parent concepts. The latter is referred to as "completed" annotation. Additional filtering can be performed to remove concepts annotated by many genes (gray letters). **(B)** Parent terms can be readily obtained by traversal of the ontology structure (a node represents a concept, an arrow an *is a* or *part of* relationship among concepts; terminal or leaf concepts are denoted in black). **(C)** Gene function prediction of the unknown gene, denoted by I, by using the majority voting approach: the annotation of all immediate neighbors in a co-expression network (black ellipses) is considered. **(D)** Deriving a prediction for the gene I by ranking the annotation obtained through its neighbors. By using the raw annotation, unambiguous prediction cannot be derived (left column); the "completed" annotation aids in deriving meaningful predictions by considering concepts intermediate in the hierarchy (e.g., concept c, middle column); additional filtering (right column) further improves the prediction ("optimized ontology").

Particularly, its reliance on the immediate neighborhood of a given node renders it applicable in estimating usefulness of local information on gene function prediction.

Here, the network consists of nodes corresponding to the genes included in the aforementioned *Arabidopsis* transcriptomics compendium. The necessary steps to transform similarity of gene expression profiles to edges between genes in a final co-expression network rely on the approach presented in Mutwil et al. (2011). In summary, this approach is comprised of ranking pair-wise gene expression profiles by the Pearson correlation coefficient. Successively, the application of statistical tests is conducted to determine the optimal cut-off (range) for the reciprocal ranks which translate into establishing edges between the nodes in the network. Moreover, an optimality principle is employed to select a set of best-performing parameter values with respect to the GBA principle. To this end, we conduct an iterative search on the allowable ranges for the reciprocal ranks that maximize the similarity of gene function in the neighborhood of a given gene/node. A highest reciprocal rank (HRR) cut-off between 10 and 30 produced biologically relevant networks (Mutwil et al., 2010). However, while >80% of the nodes were disconnected for HRR = 10, and consequently excluded from any further co-expression analysis, a HRR = 30 was chosen as the number of disconnected nodes decreased to 25%. Note, that by relying on ranks of derived from pair-wise correlations of gene expression profiles, no explicit threshold for the Pearson correlation coefficient is needed. This is nicely illustrated by the range of Pearson correlation coefficients of expression profiles of a pairs of genes with a HRR of 30 which varies from 0.32 to 0.9 depending on the individual gene. The advantage of using HHR rather than the simple pair-wise

correlation is that co-expression analysis by HRR uncovers more meaningful biological associations (Aoki et al., 2007).

The obtained co-expression network is composed of 9,994 nodes, which correspond to those genes in *Arabidopsis*' genome that are annotated with a set of concepts from all three ontologies, i.e., MapMan, GO-BP, and GO-MF. This network consists of 461 connected components of which 439 are singleton genes, i.e., nodes with no adjacent edges, and exhibits a density of 0.001. The largest component contains 9,506 nodes and the average degree of a node is 10.36. To simulate the effect on gene function prediction depending on the ontology used, the annotation provided in all three ontologies for a set of randomly chosen genes is discarded. To this end, the number of this artificially unannotated genes is set to be 4,000, a fraction corresponding to the ∼40% genes of unknown function in *Arabidopsis*.

For each of the 4,000 genes, the annotations of all adjacent nodes are derived using the completed ontology and ordered in a list, separately for all three ontologies. Every concept present in the annotation of neighboring nodes is ranked from the most to the least frequently appearing within the neighborhood (**Figure 9**). The function of an unannotated gene is then predicted by examining the first k functions in the list. Here, we consider the predictions of the top $k \in [1,20]$ most abundant concepts in the network vicinity and successively evaluate them by comparing the predicted terms to the original discarded annotation. This procedure is repeated 1,000 times, such that in every iteration a different set of randomly unannotated genes is sampled and evaluated for every k most abundant concepts.

Furthermore, we removed those 20 concepts (corresponding to the choice of parameter k) with the lowest IC from all three complete ontologies. The aim of this filtering step is to avoid deriving trivial annotation (e.g., the root concepts of the ontologies) or unspecific annotations (e.g., very broad, high-level biological concepts) as predictions. We note that although those high-level terms are technically correct in terms of prediction, their benefit in characterizing a gene of unknown function is limited (cf. **Figure 9D**). An example of terms exhibiting a low IC are within the GO-BP sub-ontology "biological process" (GO:0008150), i.e. the root term or "transport" (GO:0006810). For GO-MF, examples of removed terms include "binding" (GO:0005488) and, again, the root node "molecular function" (GO:0003674). In contrast, more specific concepts of higher IC are unaffected by this filtering step. These include, for instance, the children of the term "binding" which are "secretion" (GO:0046903) and "ion transport" (GO:0006811). These modified ontologies are termed "optimized ontologies" and further used for evaluation of the prediction

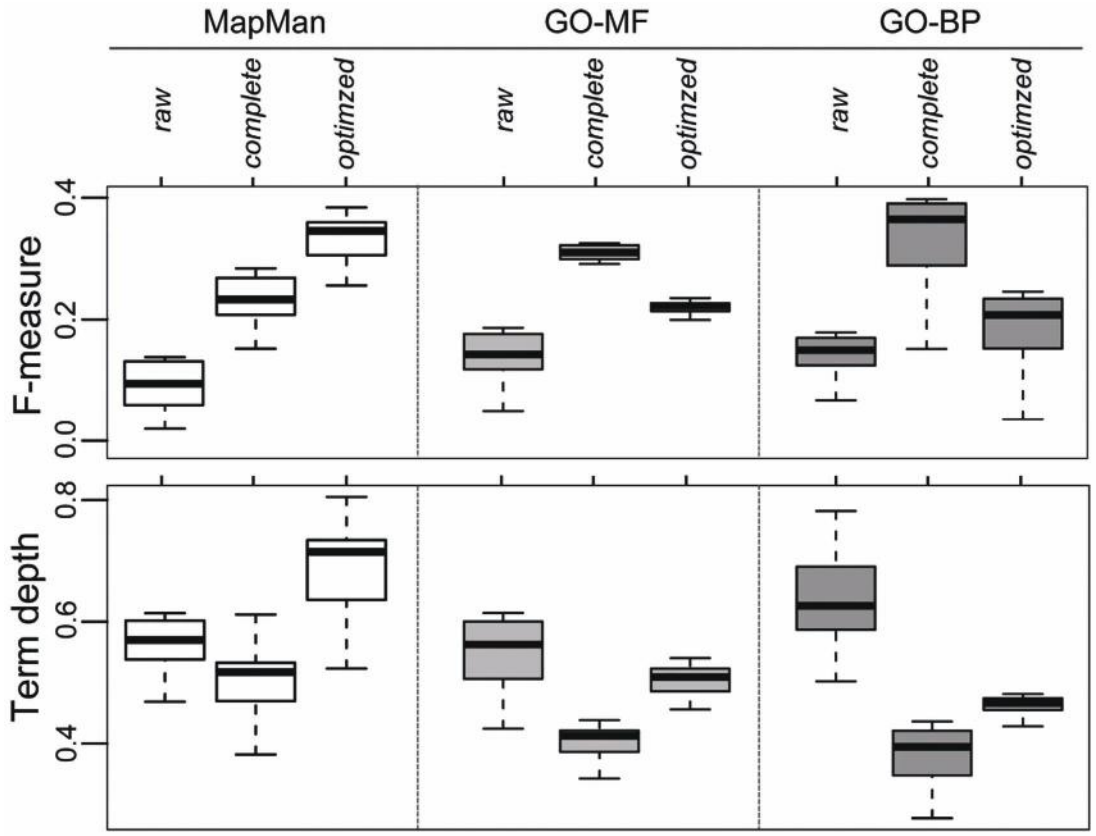


FIGURE 10 | Effect of the preprocessing of ontologies (cf. Figure 9) and the impact on network-based gene function prediction by majority voting for all three ontologies quantified by the F-measure (upper panel) and the normalized depth of term/concept (lower panel) separately for raw, complete and optimized versions of the ontologies.

performance (**Figure 8**). Finally, the effect of this optimization step on gene function prediction is illustrated in **Figure 10**: A raw ontology only contains some of the ancestral concepts resulting in a lower prediction performance (F-measure; similar results hold for precision and recall; data not shown) and average term depth of predicted concepts (similar results hold for the average IC of predicted terms; data not shown). In contrast, the complete ontology includes all ancestral concepts defined in the respective ontology, resulting in an increase of prediction performance; however, it is accompanied by a lower term depth of predicted concepts. The optimized ontology removes ambiguous terms, i.e., terms of high IC, and represents a compromise between good prediction performance and specificity of derived predictions. Interestingly, MapMan profits the most from the proposed optimization strategy.

EVALUATION OF GENE ANNOTATION PREDICTION PERFORMANCE

The quality of the predicted ontological concepts for genes is evaluated by two complementary strategies. While the first strategy comprises the use of classical quality measures from the field of pattern recognition and information retrieval that assess the correctness of predicted terms, the second strategy seeks to quantify the quality of those derived predictions in terms of biological relevance. Again, the previously established concepts of term depth and IC are employed for this task. Note that for the purpose of comparative evaluation, both term depth and IC are normalized to the respective maximum value encountered within the particular ontology.

For a single gene, the prediction performance for a set of derived concepts, C_p, is used in defining the precision as:

$$\text{precision} = \frac{|C_p \cap C_a|}{|C_p|},$$

where C_a denotes the set of originally annotated concepts. Furthermore, we define the recall of the prediction of concepts for the gene as:

$$\text{recall} = \frac{|C_p \cap C_a|}{|C_a|}.$$

Finally, we rely on the F-measure as a combined performance index, defined as the harmonic mean of precision and recall:

$$F = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}.$$

In the case of precision = recall = 0, we take $F = 0$. Note, that the values for all three performance indices correspond to the average of precision, recall, and F-measure, respectively, for all artificially unannotated genes over 1,000 iterations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Bioinformatics_and_Computational_Biology/10.3389/fgene.2012.00115/abstract

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y.-H. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Miklos, X., Abril, J. F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. D., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J, Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage,T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.

Aoki, K., Ogata, Y., and Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.* 48, 381–390.

Ashburner, M. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press, Addison-Wesley.

Bard, J. B. L., and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* 5, 213–222.

Blake, J. A., and Harris, M. A. (2002). "The gene ontology (go) project: structured vocabularies for molecular biology and their application to genome and expression analysis," in *Current Protocols in Bioinformatics*, Chap. 23, ed. R. D. M. Page (Hoboken: John Wiley & Sons, Inc.), 7.2.1–7.2.9.

Bondy, J. A., and Murty, U. (2008). *Graph Theory*. New York: Springer.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Hub, T. A., and Group, T. W. P. W. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288–289.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., and Wong, E. D. (2012). *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, D700–D705.

Doehlemann, G., Wahl, R., Horst, R. J., Voll, L. M., Usadel, B., Poree, F., Stitt, M., Pons-Kühnemann, J., Sonnewald, U., Kahmann, R., and Kämper, J. (2008). Reprogramming a maize plant: transcriptional and metabolic changes induced by the fungal biotroph *Ustilago maydis*. *Plant J.* 56, 181–195.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210.

Goeman, J. J., and Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 24, 537–544.

Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M. (2011). Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief. Bioinformatics*. doi: 10.1093/bib/bbr066. [Epub ahead of print].

Harris, M. A., and Gene Ontology, C. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). Array quality metrics – a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416.

Klie, S., Nikoloski, Z., and Selbig, J. (2010). Biological cluster evaluation for gene function prediction. *J. Comput. Biol.* 17, 1–18.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210.

Lehninger, A., Nelson, D., and Cox, M. (2008). *Lehninger Principles of Biochemistry*. New York: W. H. Freeman.

Leskovac, V. (2003). *Comprehensive Enzyme Kinetics*. New York: Springer.

Mochida, K., Uehara-Yamaguchi, Y., Yoshida, T., Sakurai, T., and Shinozaki, K. (2011). Global landscape of a co-expressed gene network in barley and its application to gene discovery in *Triticeae* crops. *Plant Cell Physiol.* 52, 785–803.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z., and Persson, S. (2011). Planet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910.

Mutwil, M., Usadel, B., Schutte, M., Loraine, A., Ebenhoh, O., and Persson, S. (2010). Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant Physiol.* 152, 29–43.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. (2009). ATTED-II provides coexpressed gene networks for *Arabidopsis*. *Nucleic Acids Res.* 37, D987–D991.

Obayashi, T., and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.* 16, 249–260.

Punta, M., and Ofran, Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput. Biol.* 4, e1000160. doi:10.1371/journal.pcbi.1000160

Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Los Altos, 448–453.

Rivals, I., Personnaz, L., Taing, L., and Potier, M. C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 23, 401–407.

Rotter, A., Usadel, B., Baebler, S., Stitt, M., and Gruden, K. (2007). Adaptation of the mapman ontology to biotic stress responses: application in *Solanaceous* species. *Plant Methods* 3, 10.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., and Mewes, H. W. (2004). The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* 32, 5539–5545.

Schiwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in Yeast. *Nat. Biotechnol.* 18, 1257–1261.

Sokal, R. R., and Rohlf, F. J. (2003). *Biometry*. New York: W.H. Freeman and Company.

Stein, L. D. (2003). Integrating biological databases. *Nat. Rev. Genet.* 4, 337–345.

Stevens, R., Goble, C. A., and Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Brief. Bioinformatics* 1, 398–414.

Tellström, V., Usadel, B., Thimm, O., Stitt, M., Küster, H., and Niehaus, K. (2007). The lipopolysaccharide of *Sinorhizobium meliloti* suppresses defense-associated gene expression in cell cultures of the host plant *Medicago truncatula*. *Plant Physiol.* 143, 825–837.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, A. S. Y., and Stitt, M. (2004). Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939.

Tsiaras, V., Triantafilou, S., and Tollis, I. (2008). "Treemaps for directed acyclic graphs graph drawing," eds S.-H. Hong, T. Nishizeki and W. Quan (Heidelberg: Springer), 377–388.

Urbanczyk-Wochniak, E., Usadel, B., Thimm, O., Nunes-Nesi, A., Carrari, F., Davy, M., Bläsing, O., Kowalczyk, M., Weicht, D., Polinceusz, A., Meyer, S., Stitt, M., and Fernie, A. (2006). Conversion of mapman to allow the analysis of transcript data from & lt; I & gt; *Solanaceous* species: effects of geznetic and environmental alterations in energy metabolism in the leaf. *Plant Mol. Biol.* 60, 773–792.

Yon Rhee, S., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509–515.

Zeeberg, B., Feng, W., Wang, G., Wang, M., Fojo, A., Sunshine, M., Narasimhan, S., Kane, D., Reinhold, W., Lababidi, S., Bussey, K., Riss, J., Barrett, J., and Weinstein, J. (2003). Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4, R28.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 April 2012; paper pending published: 03 May 2012; accepted: 05 June 2012; published online: 28 June 2012.

Citation: Klie S and Nikoloski Z (2012) The choice between MapMan and Gene Ontology for automated gene function prediction in plant science. Front. Gene. ***3****:115. doi: 10.3389/fgene.2012.00115*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

METHODS ARTICLE
published: 26 April 2013
doi: 10.3389/fgene.2013.00062

Use of the protein ontology for multi-faceted analysis of biological processes: a case study of the spindle checkpoint

Karen E. Ross[1], Cecilia N. Arighi[1], Jia Ren[1], Darren A. Natale[2], Hongzhan Huang[1] and Cathy H. Wu[1,2]*

[1] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
[2] Protein Informatics Resource, Georgetown University Medical Center, Washington, DC, USA

Edited by:
John Hancock, University of Cambridge, UK

Reviewed by:
Rafael D. Mesquita, Universidade Federal do Rio de Janeiro, Brazil
Dong Xu, Idaho State University, USA

***Correspondence:**
Cathy H. Wu, Center for Bioinformatics and Computational Biology, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA.
e-mail: wuc@dbi.udel.edu

As a member of the Open Biomedical Ontologies (OBO) foundry, the Protein Ontology (PRO) provides an ontological representation of protein forms and complexes and their relationships. Annotations in PRO can be assigned to individual protein forms and complexes, each distinguishable down to the level of post-translational modification, thereby allowing for a more precise depiction of protein function than is possible with annotations to the gene as a whole. Moreover, PRO is fully interoperable with other OBO ontologies and integrates knowledge from other protein-centric resources such as UniProt and Reactome. Here we demonstrate the value of the PRO framework in the investigation of the spindle checkpoint, a highly conserved biological process that relies extensively on protein modification and protein complex formation. The spindle checkpoint maintains genomic integrity by monitoring the attachment of chromosomes to spindle microtubules and delaying cell cycle progression until the spindle is fully assembled. Using PRO in conjunction with other bioinformatics tools, we explored the cross-species conservation of spindle checkpoint proteins, including phosphorylated forms and complexes; studied the impact of phosphorylation on spindle checkpoint function; and examined the interactions of spindle checkpoint proteins with the kinetochore, the site of checkpoint activation. Our approach can be generalized to any biological process of interest.

Keywords: protein ontology, biocuration, phosphorylation, spindle checkpoint, kinetochore

INTRODUCTION

Understanding the meaning of data is essential for accurate scientific analysis and interpretation. Ontologies formalize the meaning of terms using a defined vocabulary that facilitates the integration of data and knowledge (Gkoutos et al., 2012). Interoperability of ontological resources is required to automatically analyze data across different data repositories and to enable automatic reasoning for knowledge discovery (Hoehndorf et al., 2011). The Open Biological and Biomedical Ontologies (OBO) Foundry is a collaborative initiative[1] whose goal is to create and maintain an evolving collection of non-overlapping interoperable ontologies that will offer unambiguous representations of the types of entities in biological and biomedical reality (Ceusters and Smith, 2010). The OBO Foundry establishes best ontology practices, including adoption of a common formal language, high standards for documentation, and collaborative development (Smith et al., 2007).

Within the Foundry, the Protein Ontology (PRO[2]) is charged with the formal representation of protein-related classes (Natale et al., 2011). PRO has three sub-ontologies informally referred to as ProEvo, ProForm, and ProComp. Classes in ProEvo represent proteins that are evolutionarily related based on full-length sequence similarity. Classes in ProForm include species-specific and species-independent classes of protein isoforms, co- and post-translationally modified (PTM) forms, and variant forms. Finally, classes in ProComp encompass protein-containing complexes with formal descriptions of their components, facilitating robust annotation of variations in composition and function contexts for protein complexes within and between species (Bult et al., 2011).

Protein Ontology terms are labeled with categories to reflect their position in the PRO hierarchy. These categories are: (i) family: protein products of a distinct gene family arising from a common ancestor; (ii) gene: the protein products of a distinct gene; (iii) sequence: protein products that have a distinct sequence upon initial translation; and (iv) modification: protein products derived from a single mRNA species that differ because of some change (or lack thereof) that occurs after the initiation of translation (co- or post-translational; Natale et al., 2011).

To facilitate reliable communication and management of data, PRO is organized under the umbrella of the Basic Formal Ontology (BFO), a top-level formal foundational ontology in the biomedical domain. BFO represents, in consistent fashion, the upper level categories common to ontologies developed in different domains and at different levels of granularity. It adopts a view of reality as comprising (1) continuants: entities that continue or persist through time (objects, qualities, and functions), and (2) occurrents: the events or happenings in which continuants participate[3]. In this schema, PRO falls under continuants (object) at the molecule level.

[1] http://www.obofoundry.org/
[2] http://www.proconsortium.org/
[3] http://precedings.nature.com/documents/1941/version/1/files/npre20081941-1.pdf

The relations used in PRO are defined in the OBO Relation Ontology (Smith et al., 2005), an ontology commonly used among the OBO Foundry ontologies.

Moreover, PRO interoperates seamlessly with other OBO ontologies by reusing terms whenever the classes needed already exist in other ontologies. This is the case for the protein complex terms found in the Cellular Component branch of the Gene Ontology (GO; Ashburner et al., 2000), which provides the species-independent protein complex terms for PRO. Therefore, most of the terms in ProComp are children of GO terms. Similarly, other ontologies are used for the logical definition of PRO terms. In particular, the Protein Modification Ontology (PSI-MOD; Montecchi-Palazzi et al., 2008) is used for amino acid residue modification terms, and NCBI taxonomy[4] is used for species terms.

In addition, PRO leverages and cross references data in existing protein-centric informatics resources. For example, UniProtKB (Bult et al., 2011) is the main source for species-specific protein and isoform terms, and Reactome (Croft et al., 2011) is the main source for human protein complexes and protein modified forms. In this way, PRO offers the ontological representation for the entries in these resources, facilitating data integration.

The formal definition of protein forms and complexes at various levels of granularity in the PRO framework provides a means to associate annotations to the most appropriate class, as opposed to the traditional gene-level-only association. This is especially useful, for example, in cases where functions are realized by protein complexes rather than their individual components, or by specific isoforms of a protein, or by a protein modified form. Class-specific annotations are stored in PRO using controlled vocabularies and are integrated in the PRO website so they can be searched. Therefore, the PRO framework, along with the annotation and the mapping to relevant bioinformatics resources help to answer biologically important questions, such as: (1) What proteins and complexes are involved in a particular process? (2) What proteins and complexes are conserved in a given set of species? and (3) What function(s) is associated with a given protein form or complex?

To be able to answer the questions described in the previous section, PRO has to provide an adequate coverage of terms and annotations that pertain to the biological questions being asked. The ultimate goal in PRO is the representation of protein-related terms for the 12 GO Reference Genomes and human protein complexes from Reactome. Release 32.0 contains 35,196 PRO terms from which about 25,000 are ProEvo terms (family and gene-level classes), 9,500 are ProForm terms (isoforms and modified forms), and 393 are ProComp terms. In terms of annotations, there are 2,941 GO annotations derived from 1,242 publications. The distribution files[5] include the ontology in OBO format (pro.obo), the accompanying annotation file (PAF.txt) in a tab-delimited format, and mappings to external databases, also tab delimited. PRO is also available in OWL format through BioPortal at the National Center for Biomedical Ontologies (NCBO; Musen et al., 2012).

In this article we use the features of PRO, including a graphical representation of the PRO hierarchy, to explore the spindle checkpoint. The spindle checkpoint monitors interactions between kinetochores and spindle microtubules during mitosis and meiosis and inhibits the onset of anaphase until all kinetochores have made correct attachments to the spindle (Zich and Hardwick, 2010; Sun and Kim, 2012). A functional spindle checkpoint is necessary for high fidelity chromosome segregation; loss of the checkpoint increases the incidence of aneuploidy, a condition associated with cancer and birth defects in humans. The spindle checkpoint is well conserved in eukaryotes and depends on seven core checkpoint proteins called BUB1, BUB1B (BubR1), AURKB (Aurora B), TTK (Mps1), MAD1L1, MAD2L1, and BUB3 in humans (Oh et al., 2010; Zich and Hardwick, 2010). The target of the checkpoint is the Anaphase-Promoting Complex/Cyclosome (APC/C), a multi-subunit ubiquitin ligase whose activity is required for the metaphase to anaphase transition. In the presence of an incomplete or defective spindle, the MCC, a protein complex consisting of the checkpoint proteins BUB1B, BUB3, and MAD2L1 and the APC/C component Cdc20 associates with the APC/C and inhibits its activity (Lara-Gonzalez et al., 2012).

The spindle checkpoint represents a rich use case with features to demonstrate the application of all three sub-ontologies of PRO. First, it has been extensively studied in a range of organisms, and the core checkpoint proteins are conserved in eukaryotes from yeast to humans. Thus, using ProEvo as a guide to the evolutionary relationships amongst spindle checkpoint proteins, it is possible to make predictions about checkpoint proteins based on evidence concerning their counterparts in other organisms. The ProEvo representation can also highlight differences between spindle checkpoint proteins that may have implications for checkpoint function. Second, the spindle checkpoint is highly dependent on phosphorylation – of the seven core spindle checkpoint proteins in vertebrates, three (BUB1, AURKB, and TTK) are confirmed protein kinases and all seven are phosphoproteins (Oh et al., 2010; Zich and Hardwick, 2010). The individual representation and annotation of modified protein forms in PRO facilitates studies of the role of phosphorylation in the checkpoint. Finally, spindle checkpoint proteins participate in numerous protein complexes, which can be captured by ProComp. Through our analysis we demonstrate that PRO can provide a logical framework to represent existing knowledge about proteins and complexes involved in a biological process and serve as a platform for making predictions for further experimental studies.

METHODS

POPULATION OF PRO WITH SPINDLE CHECKPOINT INFORMATION

Literature and data mining

Information about spindle checkpoint protein forms and their functions was identified through curation of full-length articles that were returned in a PubMed search using the keywords "Bub1," "BubR1," and "Mad3" (BubR1 is a commonly used synonym for the checkpoint protein BUB1B and MAD3 is the closest yeast relative of BUB1B). Because of our interest in phosphorylation of checkpoint proteins, we focused our curation efforts on the subset of articles that were flagged by the text mining tool Rule-based LIterature Mining System for Protein Phosphorylation (RLIMS-P)

[4] http://www.obofoundry.org/cgi-bin/detail.cgi?id=ncbi_taxonomy
[5] ftp://ftp.pir.georgetown.edu/databases/ontology/pro_obo/

as containing mentions of phosphorylation in the abstract (Yuan et al., 2006). We extracted information on all proteins for which there was experimental data in the articles we curated, thereby expanding our analysis of the checkpoint beyond the three proteins we used as keywords for the PubMed search. In addition, we mined three curated interaction databases [Molecular INTeraction Database (MINT[6]; Chatr-Aryamontri et al., 2007; release date 10/26/2012); IntAct[7] (Kerrien et al., 2012; release 159); and the Biological General Repository for Interaction Datasets (BioGRID[8]; Stark et al., 2011; release 3.1.94)] for all direct physical interactions that had been demonstrated in low throughput experiments involving proteins identified in our literature search.

RACE-PRO: PRO community annotation interface

All information on protein forms was entered into Rapid Annotation interfaCE for PRO (RACE-PRO[9]), a web-based interface for PRO community annotation. This interface is intended for any user independent of their ontology knowledge. It allows the specification of a protein form by entering the protein sequence and features (protein regions, and/or modified residues) with the evidence source (usually literature), and the functional annotation associated with the given protein form using controlled vocabularies, such as GO for processes, functions, and subcellular location, and Pfam[10] (Punta et al., 2012) for protein domains. Currently, RACE-PRO cannot be used for protein complex or protein family terms, although an expanded version of RACE-PRO that would enable these capabilities is under development. Instead, a user can request complex and family terms via the SourceForge PRO tracker[11]. Links to both RACE-PRO and the PRO tracker can be found on the PRO home page.

The RACE-PRO entries were checked by a PRO editor and converted to PRO terms using a semi-automated process, in which standard names and definitions for gene level and isoform level terms are automatically generated as are missing parent terms that are necessary to complete the PRO hierarchy. Definitions of modified protein forms and PRO terms for complexes and families were handled manually. The end result of the processing pipeline were OBO stanzas containing the term IDs, names, definitions, synonyms, categories, and relationships to other terms. Annotations were included in the PRO Annotation File (PAF). All terms and annotations generated in this study can be found in PRO release 32.

ANALYSIS AND VISUALIZATION OF THE PRO TERMS

Once data was entered into the PRO framework, it was analyzed and visualized using the search and graphical display tools in the PRO website. The search functionality allows all parts of a PRO entry, including definition and annotation, to be searched. Query terms can be words or phrases or unique identifiers from other resources such as Pfam or GO. Searches can be restricted to a particular field of a PRO entry; for example, searching for the term "9606" in the Taxon ID field will retrieve all human protein terms. The search terms "NOT NULL" and "NULL" can be used to identify PRO entries that do or do not contain information in a selected field. Multiple search terms can be joined with the Boolean terms "AND," "OR," and "NOT" to carry out more complex searches. In addition, searches can be restricted to particular categories of PRO entries such as modified forms, disease-related forms, or complexes using the "Quick Links" menu provided on the PRO search page. Finally, the search result table can be customized to include/remove information and can be downloaded in tab-delimited format.

The PRO hierarchy can be visualized using a built-in tool based on Cytoscape Web (Lopes et al., 2010). The tool can be accessed by clicking on the "Cytoscape view" icon on any PRO entry page. The display can be set to show the parent(s), siblings, and/or children of the entry with or without organism-specific terms. Either sequence level or modification-level child terms can be viewed. Advanced display options allow the user to show or hide nodes based on their PRO Category (e.g., "organism-gene" or "complex") and to hide individual nodes of choice. Selecting any node in the display provides the option to jump to the Cytoscape web view, PRO entry page, or text-based hierarchy for that node. Using the batch entry mode, the user can add terms to the display by entering their PRO or GO IDs as a comma separated list. A feature that displays the Cytoscape Web view of multiple terms selected from the PRO search results page will be available soon.

ANALYSIS OF PRO DATA WITH EXTERNAL TOOLS

The kinetochore protein–protein interaction (PPI) network was displayed using locally installed Cytoscape, version 2.8 (Smoot et al., 2011). To construct the network, we first searched PRO for all terms annotated with kinetochore or centromere localization using the query: "Taxon ID 9606 (human) AND Ontology ID GO:0000776 (kinetochore) OR Taxon ID 9606 (human) AND Ontology ID GO:0000779 (condensed chromosome, centromeric region)," and downloaded the OBO stanzas and PAF for the 34 search results. Using a script (available upon request), we extracted the name, definition, category, and label (PRO-short-label) from the OBO stanzas as well as parent-child and kinase-substrate relationships. Parent-child relationships (identified by the "is_a" relation) were directly extracted from the PRO terms. Kinase information appears in the free-text comment field of the OBO stanza; however, it could be parsed out because it is entered by PRO curators in a standardized format (Kinase = "name"; PRO ID). Protein binding related annotations (identified by the GO evidence code "inferred from physical interaction" or IPI) were extracted from the PAF. The script then generated two tab-delimitated text files, which are importable into Cytoscape: a network file containing each pair of interacting proteins, its interaction type, and corresponding evidence and a PRO entry information file containing PRO ID and entity description. Those two files were further converted into visualized protein networks with the Cytoscape functions "Import → Network from table" and "Import → Attribute from table" functions. In these networks, each node is a PRO entry and two nodes were connected by an edge if they were associated by a relation. Entity descriptions and relations annotations were represented as node or edge attributes.

[6]http://mint.bio.uniroma2.it/mint/
[7]http://www.ebi.ac.uk/intact/
[8]http://www.thebiogrid.org
[9]http://pir.georgetown.edu/cgi-bin/pro/race_pro
[10]http://pfam.sanger.ac.uk/
[11]http://sourceforge.net/tracker/?group_id=266825&atid=1135711

Multiple sequence alignments were performed using ClustalW version 2.1 (Larkin et al., 2007; Goujon et al., 2010) and visualized with Jalview Desktop version 2.8 (Waterhouse et al., 2009). Experimentally determined phosphorylation sites taken from PRO phosphorylation site data and phosphorylation sites predicted based on sequence alignment were highlighted in the Jalview display.

RESULTS

OVERVIEW OF THE PRO REPRESENTATION OF THE SPINDLE CHECKPOINT

To get an overview of the extent of spindle checkpoint-related information contained within PRO we performed a search in PRO for terms containing the phrases "spindle checkpoint," "spindle assembly checkpoint," or "mitotic checkpoint." The search returned 112 PRO terms. The PRO search query and the Cytoscape web view of the combined hierarchy of the search result terms are shown in **Figure 1**. The hierarchy, which includes parents and children of the search result terms as well as complexes containing the search result terms, consists of 208 terms (including two obsolete terms) spanning all levels in PRO. There are three family level terms – Histone H2A (PR:000027547), Aurora Kinase (PR:000035365), and BUB1/BUB1B (PR:000035665) – and 21 gene-level terms, including the seven core checkpoint proteins. Of the 35 modification-level terms, 26 are phosphorylated forms, 6 are unphosphorylated forms, 1 is an acetylated form, and 1 is a cleaved form. The figure also includes one sequence level term

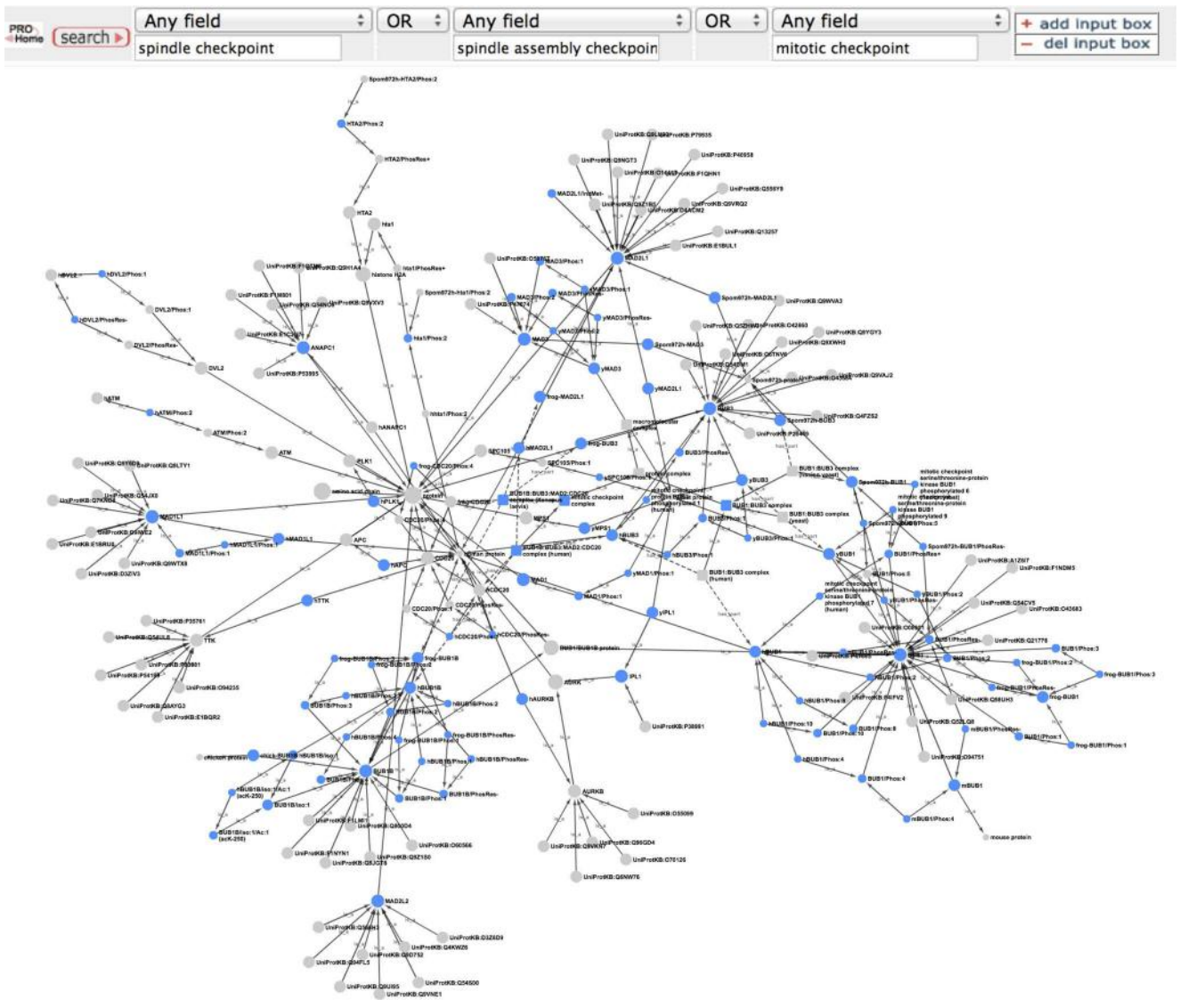


FIGURE 1 | Overview of the PRO representation of the spindle checkpoint. PRO search query to retrieve PRO terms that contain the phrases "spindle checkpoint" or "spindle assembly checkpoint" or "mitotic checkpoint" and combined Cytoscape web view of the search results. In the Cytoscape view, nodes retrieved by the search are blue; related nodes (parents and children) are gray. Circles represent proteins or protein forms; squares represent protein complexes.

[BUB1B isoform 1 (PR:000028795)]; two complexes [BUB1:BUB3 complex (PR:000035566) and the mitotic checkpoint complex (MCC; GO:0033597)]; and the high level terms amino acid chain (PR:000018263), protein (PR:000000001), macromolecular complex (GO:0032991), and protein complex (GO:0043234). The 140 organism-specific terms (75 organism-gene, 47 organism-modification, 1 organism-sequence, and 17 organism-complex terms) span a wide evolutionary range, including terms from humans, rodents, frogs, plants, insects, worms, and yeast. We will consider some specific questions that can be addressed by this representation in the sections that follow.

EVOLUTIONARY RELATIONSHIP OF BUB1, BUB1B, AND MAD3

The spindle checkpoint pathway is highly conserved throughout eukaryotes. Homologs of the core checkpoint proteins are present in organisms from yeast to humans and checkpoint mechanisms, such as MCC inhibition of the APC/C, are also conserved (Zich and Hardwick, 2010; Lara-Gonzalez et al., 2012). Despite the overall similarity, there are significant differences in the details of the sequence and function of some the checkpoint proteins. One of the most striking examples of this variation involves the "BUB-like" proteins, BUB1, BUB1B, and MAD3. Derived from a common ancestor, modern BUB-like proteins arose as the result of multiple gene duplication events. Some organisms have only one of these proteins; others, like *Arabidopsis thaliana*, have as many as three (Suijkerbuijk et al., 2012). Humans have two (BUB1 and BUB1B). Budding and fission yeasts also have two: BUB1, which is orthologous to human BUB1, and MAD3, which is most closely related to human BUB1B. BUB1, BUB1B, and MAD3 share an N-terminal domain containing tetratricopeptide repeats [TPR domain; (D'Arcy et al., 2010)]. This domain of budding yeast MAD3 has been shown to bind to the APC/C subunit, CDC20, an interaction critical for checkpoint-mediated inhibition of anaphase onset (Hardwick et al., 2000). Outside of this N-terminal region, however, BUB1, BUB1B, and MAD3 diverge significantly. BUB1 and BUB1B contain a C-terminal kinase domain, which is absent from MAD3. BUB1 is a bona fide protein kinase, whereas BUB1B is likely to be a pseudokinase, although BUB1B kinase activity, particularly auto-phosphorylation activity under some conditions, remains a possibility (Guo et al., 2012; Suijkerbuijk et al., 2012).

What can we learn about the evolutionary relationship of BUB1, BUB1B, and MAD3 using the PRO website?

In PRO, ProEvo classes provide insight into the evolutionary relationships among proteins by grouping proteins that share full-length sequence similarity. Importantly, this higher level relationship based on a common domain organization can be searched in PRO, as terms in ProEvo are annotated with domain information from resources such as Pfam. Therefore, we searched PRO for proteins that contained the conserved N-terminal TPR domain found in all of the BUB-like proteins (PFAM:PF08311, MAD3/Bub1 homology domain I). The search returned two results: the MAD3 gene-level term (PR:000035499) and the BUB1/BUB1B family level term (PR:000035665).

To reveal the common and divergent attributes of these protein classes, the result table was customized, via the Display Option functionality, to display the corresponding annotations and allow their direct comparison (**Figure 2A**). As expected both groups are annotated as containing the MAD3/Bub1 homology domain I (PFAM:PF08311), and the definition of the BUB1/BUB1B family states in part that: "Members of this class are related to MAD3." However, the BUB1/BUB1B proteins contain a second conserved domain, the C-terminal protein kinase domain (PFAM:PF00069) that is absent in the MAD3 class.

The combined Cytoscape web view for BUB1/BUB1B and MAD3 terms is shown in **Figure 2B**. BUB1/BUB1B and MAD3 (blue nodes) are connected by the parent term "protein." BUB1 (PR:000004854) and BUB1B (PR:000004855) are both children of the BUB1/BUB1B class, indicating that these two proteins share full-length sequence similarity. BUB1 is very highly conserved with 11 organism-specific child terms ranging from yeast to human. Compared to BUB1, BUB1B is less conserved. Its children include human and frog BUB1B terms but no yeast terms. Instead, the closest yeast relative of BUB1B is MAD3 (PR:000035499).

PREDICTION OF BUB1B PHOSPHORYLATION SITES

Phosphorylation is a major mechanism of regulation in the spindle checkpoint pathway and the interplay among the checkpoint-related phosphorylation events is complex (Zich and Hardwick, 2010). There are multiple spindle checkpoint kinases, each of which has multiple substrates. Some checkpoint proteins are targeted by more than one kinase and exist in several phosphorylated forms. One such protein, BUB1B, has at least four different mitotic phosphorylated forms (Elowe et al., 2007, 2010; Matsumura et al., 2007; Wong and Fang, 2007; Huang et al., 2008; Guo et al., 2012). Phosphorylated forms of BUB1B first appear during pro-metaphase as condensed chromosomes begin to make attachments to spindle microtubules and persist until all chromosomes have made correct bipolar attachments to the spindle at metaphase. Although BUB1B was first characterized as a spindle checkpoint protein, phosphorylated forms of BUB1B have been shown to participate in spindle assembly as well.

How can we look at the different phosphorylated forms of BUB1B in PRO? Are these forms conserved and what predictions can we make?

To view the phosphorylated BUB1B protein forms in PRO, we searched for "bub1 beta" in the PRO Name field, restricting the search to phosphorylated forms using the Quick Links menu. Eleven search results were returned: four species-independent modification-level terms and seven species-specific terms. The combined Cytoscape web view of the four species-independent terms (PR:000035361, PR:000035427, PR:000035431, and PR:000035434) is shown in **Figure 3**. The four phosphorylated forms have the species-independent BUB1B gene-level term as their common parent. Each form also has one or more organism-specific children. Alongside each form is a portion of a sequence alignment of human, frog, and mouse BUB1B with experimentally confirmed form-specific phosphorylation sites highlighted in blue and predicted phosphorylation sites highlighted in red.

A

PRO Home search | Any field | AND | Any field | + add input box | − del input box

PFAM:PF08311

Display Options

2 entries | 1 page | 50 / page |

Show selected: Hierarchy Show selected: OBO / PAF OR related: OBO / PAF

PRO ID	PRO Name	PRO Term Definition	Category	Annotation						
				Modifier	Relation	Ontology ID	Ontology Term	Relative To	Interaction With	Evidence Source
PR:000035499	spindle assembly checkpoint component MAD3	A protein that is a translation product of the Saccharomyces cerevisiae MAD3 gene or a 1:1 ortholog thereof. MAD3 is related to BUB1 and BUB1B, all of which contain the Mad3/BUB1 homology region 1 (PF00831), but MAD3 lacks the C-terminal Protein kinase domain (PF00069). [PRO:KER, PMID:15970700, PMID:22698286]	gene		has_part	Pfam:PF08311	Mad3/BUB1 homology region 1			PMID:22698286
PR:000035665	BUB1/BUB1B protein	A protein that consists of a core domain composition consisting of a the Mad3/BUB1 homology region 1 (PF00831) and a C-terminal Protein kinase domain (PF00069). Members of this class are related to MAD3. [PRO:CNA, PMID:22698286]	family		has_part	Pfam:PF08311	Mad3/BUB1 homology region 1			PMID:22698286
					has_part	Pfam:PF00069	Protein kinase domain			PMID:22698286

B

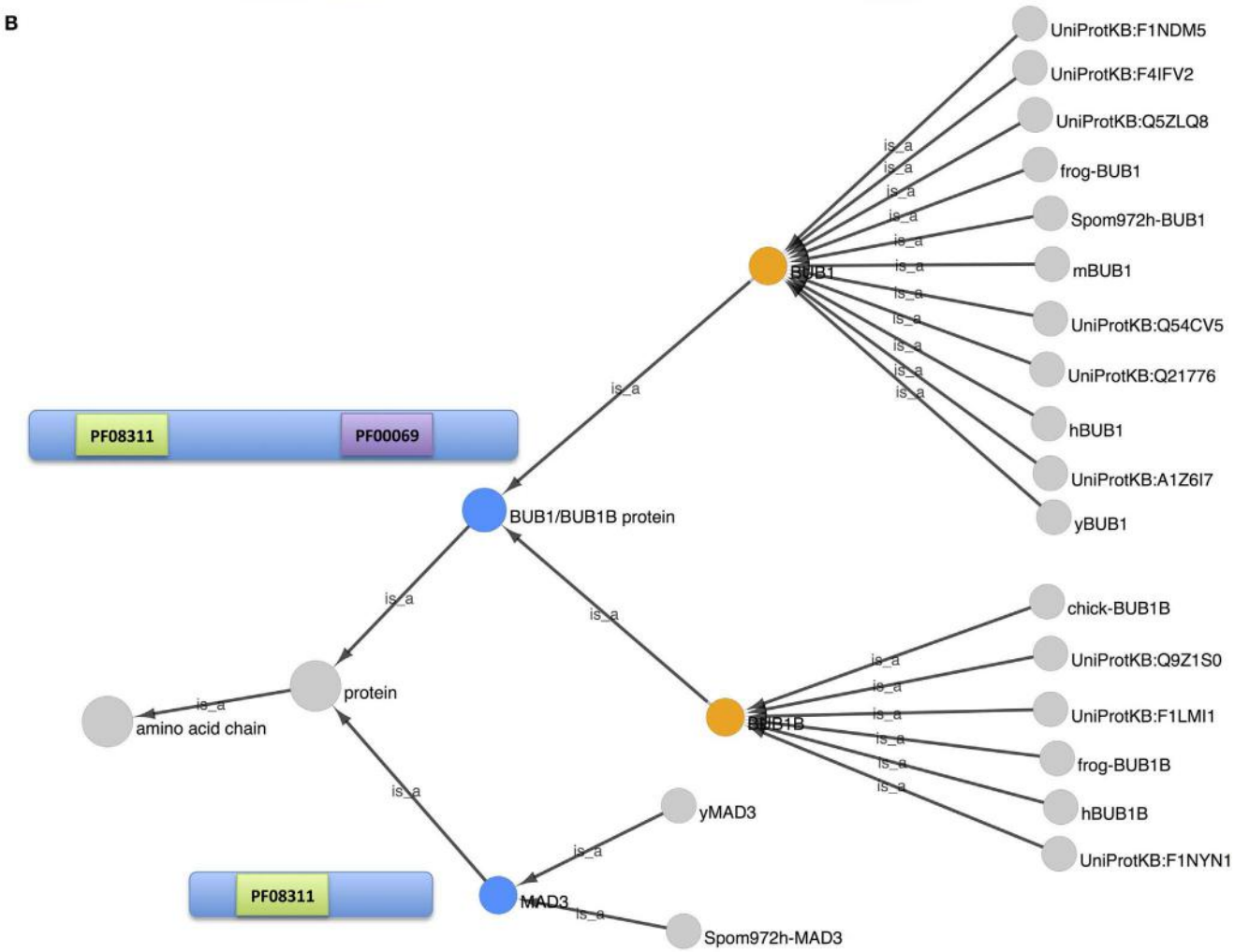


FIGURE 2 | Evolutionary relationship of BUB1, BUB1B, and MAD3. **(A)** PRO search results page showing the terms retrieved in a search for "PFAM:PF08311." Fields shown in the display were set using "Display Options." **(B)** Combined Cytoscape web view of the BUB1/BUB1B family term (PR:000035665) and the MAD3 gene-level term (PR:000035499) (blue nodes). Advanced display options were set to include family, gene, and gene-organism terms only. The BUB1 and BUB1B gene-level nodes are orange. Diagrams of the MAD3 and BUB1/BUB1B proteins show the location of Pfam domains PF08311 (MAD3/Bub1 homology domain I) and PF00069 (protein kinase domain).

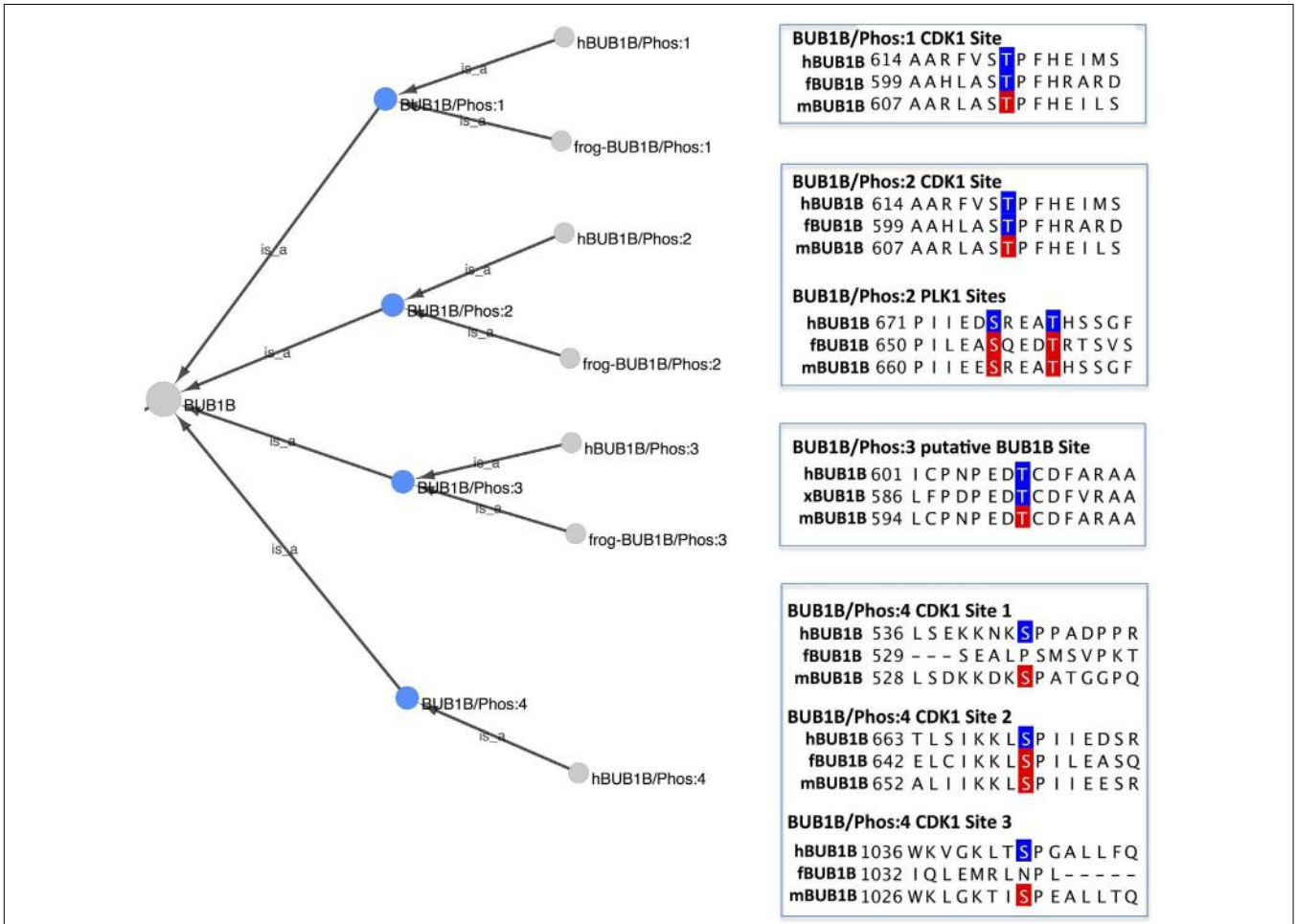


FIGURE 3 | Phosphorylated forms of BUB1B. Combined Cytoscape web view of the four species-independent BUB1B phosphorylated forms (blue nodes): BUB1B/Phos:1 (PR:000035361), BUB1B/Phos:2 (PR:000035427), BUB1B/Phos:3 (PR:000035431), and BUB1B/Phos:4 (PR:000035434). Display options were set to show parents and all children, including organism level terms. Portions of a sequence alignment of human, frog, and mouse BUB1B are highlighted to indicate experimentally determined phosphorylation sites (blue) and predicted phosphorylation sites (red).

BUB1B/Phos:1 (PR:000035361), defined in PRO as a BUB1B form that has been phosphorylated on a site analogous to Thr-620 of human BUB1B, is found in humans (PR:000035362) and frogs (PR:000035426). The frog form is phosphorylated on Thr-605, which is considered to be analogous to human Thr-620 because it aligns with human Thr-620 in a multiple sequence alignment (**Figure 3**, BUB1B/Phos:1, blue residues). In both organisms, the phosphorylation is carried out by the cyclin-dependent kinase CDK1 (see PRO entry pages, comment section).

Although BUB1B/Phos:1 has not as yet been characterized in mice, the equivalent phosphorylation site (Thr-613) is conserved in the mouse protein (**Figure 3**, BUB1B/Phos:1, red residue). Furthermore, Thr-613 of mouse BUB1B was identified as an *in vivo* phosphorylation site in a high throughput study of mitotic phosphorylation (Hegemann et al., 2011). Thus, there is a high probability that BUB1B/Phos:1 exists in mice as well.

BUB1B/Phos:2 (PR:000035427) contains the same CDK1 phosphorylation site (Thr-620 in humans) as BUB1B/Phos:1 and is additionally phosphorylated on several sites by PLK1/PLX1. Because experimental evidence indicates that PLK1 phosphorylation of BUB1B is low in the absence of prior CDK1 phosphorylation, PRO does not have a term for BUB1B phosphorylated by PLK1 alone (Elowe et al., 2007; Wong and Fang, 2007). As described in its PRO definition, human BUB1B/Phos:2 (PR:000035428) is observed during pro-metaphase when kinetochores are undergoing attachment to the mitotic spindle and under conditions that depolymerize the spindle (nocodazole treatment) or that disrupt the ability of microtubules to apply tension across kinetochores (taxol treatment).

The PLK1 phosphorylation sites in BUB1B/Phos:2 are a subject of ongoing investigation. The PRO entry page for the human BUB1B/Phos:2 (PR:000035428) documents two neighboring sites – Ser-676 and Thr-680 – that have been verified *in vivo* and two other sites – Thr-792 and Thr-1008 – that have so far only been observed in *in vitro* studies. The *in vivo* sites are shown in the sequence alignment in **Figure 3** (BUB1B/Phos:2 PLK1 sites, blue residues).

One of the challenging aspects of the curation of PRO phosphorylated forms is determining whether a phosphorylated form that has been defined in one species also exists in other species. This challenge is exemplified by BUB1B/Phos:2. There is evidence that BUB1B/Phos:2 exists in both frogs and mice, although it has not been completely characterized in either organism. All of the human BUB1B/Phos:2 phosphorylation sites that have been confirmed *in vivo* are conserved in the frog and mouse proteins (frog: Thr-605, Ser-655, and Thr-659; mouse: Thr-613, Ser-665, and Thr-669; **Figure 3**). Moreover, a phosphorylated form of BUB1B has been observed in frogs and mice in the same conditions – the presence of unattached kinetochores – under which BUB1B/Phos:2 is observed in humans (Taylor et al., 2001; Chen, 2002). This evidence alone was determined to be insufficient to create a PRO term; however, in frogs there is additional evidence in support of the existence of BUB1B/Phos:2. First, frog BUB1B is known to be phosphorylated by CDK1 on Thr-605; this phosphorylation is analogous to the CDK1 phosphorylation site in human BUB1B/Phos:2 (Thr-620). Second, as is the case in humans, CDK1 phosphorylation of frog BUB1B at Thr-605 stimulates the further phosphorylation of BUB1B by frog PLK1 (Wong and Fang, 2007). Thus, a PRO term was created for frog BUB1B/Phos:2 (PR:000035430). We predict that BUB1B/Phos:2 is also present in mice, but more experimental work is necessary to demonstrate its existence.

BUB1B/Phos:3 (PR:000035431) is phosphorylated on Thr-608 in humans (PR:000035432) and on the equivalent site, Thr-593 in frog (PR:000035433) (**Figure 3**; BUB1B/Phos:3, blue residues). An analog of BUB1B/Phos:3 has not been characterized in mice, but the phosphorylation site is conserved (mouse Thr-601; **Figure 3**; BUB1B/Phos:3, red residue) and has been shown to be phosphorylated *in vivo* (Hegemann et al., 2011). The proposed kinase for BUB1B/Phos:3 is BUB1B itself in association with the kinetochore component, CENPE (Guo et al., 2012). However, a recent structural and functional analysis indicates that BUB1B does not have kinase activity, but is instead a pseudokinase (Suijkerbuijk et al., 2012). To reflect this uncertainty, the comment section of the PRO record for the frog and human BUB1B/Phos:3 PRO entry pages states: "One of the articles cited mentions BUBR1 (PR:000026903) as the kinase when bound to CENPE (PR:000035367)."

Finally, BUB1B/Phos:4 (PR:000035435), which has so far only been observed in humans (PR:000035435), is multiply phosphorylated by CDK1 on sites distinct from those phosphorylated in BUB1B/Phos:1 and BUB1B/Phos:2. Phosphorylation occurs *in vivo* on at least three CDK1 consensus sites: Ser-543, Ser-670, and Ser-1043 (see PR:000035435, term definition). All three sites are conserved in mouse BUB1B and two of the three (mouse Ser-535 and Ser-1033) have been shown to be phosphorylated *in vivo*, strongly suggesting that BUB1B/Phos:4 exists in mouse [**Figure 3**; BUB1B/Phos:4, red residues; (Hegemann et al., 2011)]. BUB1B/Phos:4 does not exist in frogs because only one of the phosphorylation sites (human Ser-670, frog Ser-649) is conserved (**Figure 3**; BUB1B/Phos:4, red residues). However, it is noteworthy that mutation of Ser-670 alone in the human BUB1B protein produced phenotypes nearly as severe as mutating all of the BUB1B/Phos:4 sites, indicating that Ser-670 is a critical phosphorylation site (Huang et al., 2008; Elowe et al., 2010). Thus, it is possible that frog has a BUB1B form phosphorylated on Ser-649 that plays a similar role to BUB1B/Phos:4 in humans.

By combining the PRO representation of phosphorylated forms with multiple sequence alignments, we can predict not just individual phosphorylation sites, but combinations of phosphorylation sites that are likely to occur *in vivo*. Thus, we predict that mice will have a BUB1B/Phos:1 (phosphorylated on Thr-613), a BUB1B/Phos:2 (phosphorylated on Thr-613, Ser-665 and Thr-669), a BUB1B/Phos:3 (phosphorylated on Thr-601), and a BUB1B/Phos:4 (phosphorylated on mouse Ser-535, Ser-559, and Ser-1033). Frogs probably have a BUB1B/Phos:2 (phosphorylated on Thr-605, Ser-655, and Thr-659). Due to lack of phosphorylation site conservation, frogs cannot have a BUB1B/Phos:4. It would be interesting to investigate whether this difference in BUB1B phosphorylation has any biological implications.

ANALYSIS OF SPINDLE CHECKPOINT PROTEIN COMPLEXES

In the presence of unattached or incorrectly attached kinetochores, the core spindle checkpoint proteins form multiple protein complexes that contribute to the inhibition of the APC/C and metaphase arrest (Zich and Hardwick, 2010). Representation of these complexes in PRO facilitates comparisons of complex composition and the conservation of complexes across organisms. In this study, we used PRO to address questions about the APC/C inhibitory MCC and complexes containing the checkpoint kinase BUB1.

What is the function and subunit composition of the MCC?

The MCC is one of the best-characterized spindle checkpoint complexes, and consequently, it has been described in multiple bioinformatics resources, including GO and Reactome. The PRO record for the human MCC (PR:000035511), shown in **Figure 4**, demonstrates how PRO interoperates with these other resources, augmenting the representation of the complex without unnecessarily duplicating information. First, GO provides the species-independent parent term for the complex (GO:0033597; green arrow). The GO record includes the following definition of the MCC that describes its function and composition: "A multiprotein complex that functions as a mitotic checkpoint inhibitor of the anaphase-promoting complex/cyclosome (APC/C). In budding yeast this complex consists of Mad2p, Mad3p, Bub3p, and Cdc20p, and in mammalian cells it consists of MAD2, BUBR1, BUB3, and CDC20." Complex component information is also provided by Reactome (REACT 5836; red arrow). In the PRO record, the complex components are listed in the "Hierarchical Relationship" section associated with the ontological relation "has_part" (red box). Thus, PRO provides an ontological representation of the human MCC that brings together the GO definition of the complex with species-specific component information from Reactome.

Are BUB1-containing complexes conserved across species?

The BUB1 protein plays a critical role in checkpoint signal generation. Together with BUB3, it localizes to kinetochores by binding to the kinetochore component CASC5 (KNL1/blinkin) and serves as a platform for the recruitment and activation of other checkpoint proteins, including MAD1 and BUB1B (Lara-Gonzalez et al., 2012).

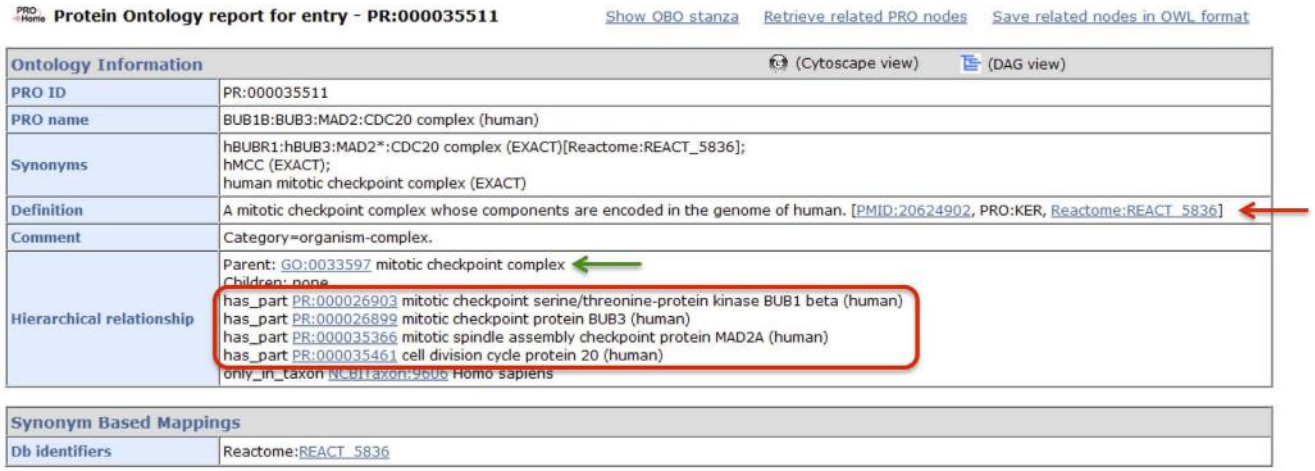


FIGURE 4 | PRO entry page for the human MCC. Screenshot of the PRO entry page for the human MCC (PR:000035511). Complex components are indicated by the red circle. Links to GO and Reactome are indicated by green and red arrows, respectively.

To view the PRO representation of BUB1-containing complexes, we searched for "BUB1" in any field and restricted the search results to complexes using the Quick Links menu. The search returned 16 results, including 11 BUB1 complexes (The other five complexes contained BUB1B rather than BUB1.). The combined Cytoscape web view of these 11 complexes and their components is shown in **Figure 5**. The complex terms (squares) and component terms (circles) that were used to generate the display are shown in blue.

BUB1 and BUB3 appear together in three different complexes: BUB1:BUB3 (PR:000035566), BUB1:BUB3:MAD1L1 (PR:000035567), and BUB1:BUB3:APC (PR:000035576) [Note: APC is the short name for the adenomatous polyposis coli protein (APC); it is not the anaphase-promoting complex/cyclosome (APC/C).]. The BUB1:BUB3 complex is highly conserved, occurring in human, fission yeast, and budding yeast (orange squares). The BUB1 proteins from all three organisms have a common parent (the species-independent BUB1 term, PR:000004854), indicating that they are orthologous; similarly, the BUB3 proteins have the species-independent BUB3 term (PR:000004856) as a common parent. Given that orthologous BUB1:BUB3 complexes exist in distantly related organisms (humans and yeast) we expect that more examples of this complex will be added to PRO in the future as more of the spindle checkpoint literature is curated. The BUB1:BUB3:MAD1L1 complex, so far observed only in humans, forms at kinetochores during the process of checkpoint activation (Seeley et al., 1999). Although the function of the BUB1:BUB3:APC complex is not known, it is interesting to note that APC, a microtubule-binding protein found at kinetochores, is phosphorylated by BUB1:BUB3 [see PRO annotation for APC (PR:000030190) and APC/Phos:1 (PR:000030182)].

The remaining BUB1-containing complexes in **Figure 5**, BUB1:BUB1B and BUB1:PLK1, illustrate the ability of PRO to represent information about the modification state of complex components. In the case of the BUB1:BUB1B complex, BUB1 can bind to unphosphorylated BUB1B, but complex formation is enhanced by the mitotic phosphorylation of BUB1B (Taylor et al., 2001). Thus, there are two human BUB1:BUB1B complexes in PRO: one consists of BUB1 and the unphosphorylated form of BUB1B (PR:000035579) and the other consists of BUB1 and BUB1B/Phos:2 (PR:000035577). The two complexes have the subunit BUB1 in common but contain different forms of BUB1B. Both complexes are children of the species-independent BUB1:BUB1B complex (PR:000035578). In the case of the BUB1:PLK1 complex, phosphorylation of human BUB1 on Ser-593 and Thr-609 by CDK1 is required for its binding to the polo-like kinase, PLK1, and for the recruitment of PLK1 to kinetochores (Qi et al., 2006). Thus, the BUB1-PLK1 complex term in PRO (PR:000035580) has only one child, the phosphoBUB1:PLK1 complex (PR:000035581) that consists of PLK1 and the CDK1-phosphorylated form of BUB1, BUB1/Phos:7.

A PROTEIN INTERACTION NETWORK FOR SPINDLE CHECKPOINT PROTEINS AT THE KINETOCHORE

The kinetochore, a complex, multi-protein structure organized around the centromeric DNA of each sister chromatid pair, is critically important as a staging area for the generation and amplification of spindle checkpoint signals (Lara-Gonzalez et al., 2012). In addition to its role in the spindle checkpoint, the kinetochore has other vital functions, including spindle microtubule binding and regulation of sister chromatid cohesion (Hori and Fukagawa, 2012). Using information downloaded from PRO, we created a network that illustrates the PPIs between checkpoint proteins and other proteins that reside at the kinetochore.

What are the PPIs observed between checkpoint proteins and other proteins in the kinetochore?

To create a PPI network of kinetochore-localized proteins, we first identified all human kinetochore-localized protein forms in PRO by searching for terms with Taxon ID 9606 (human) and

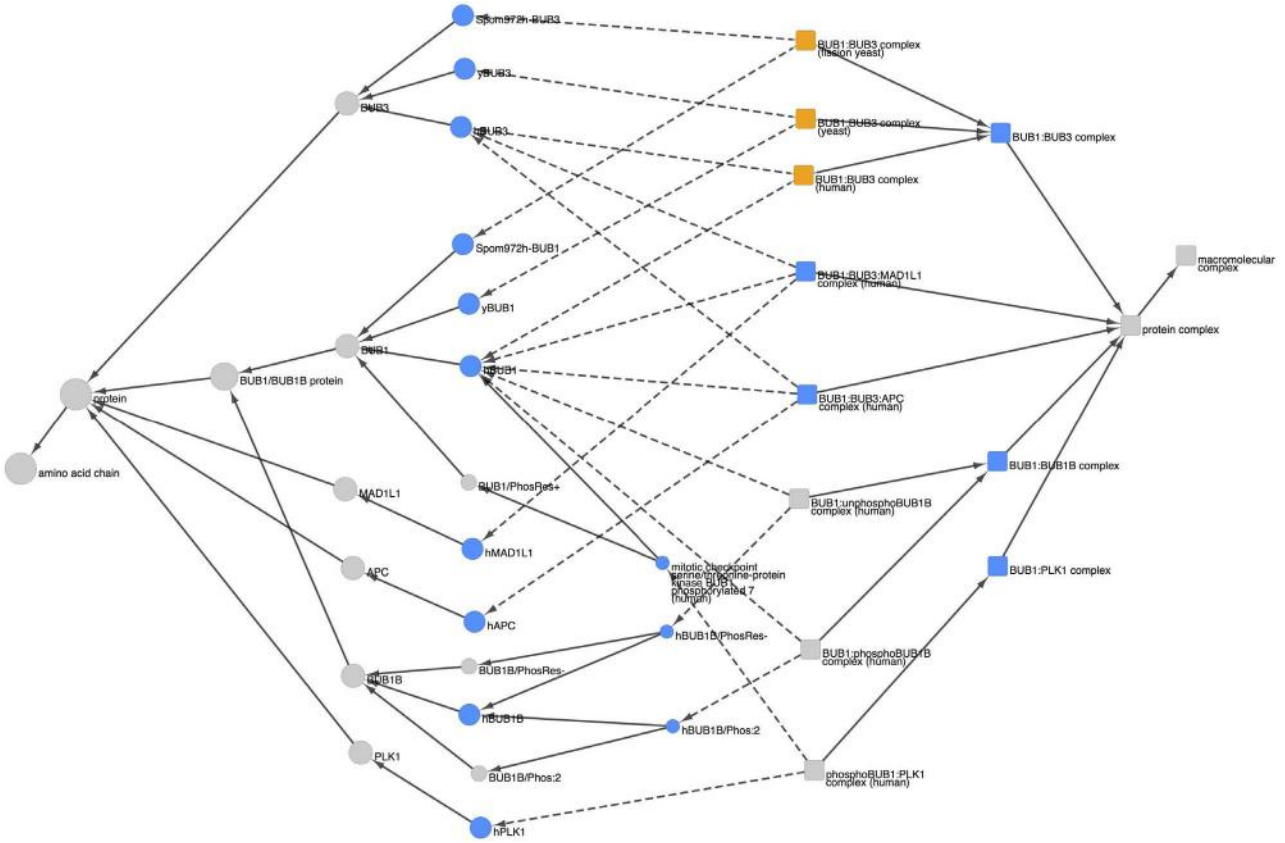


FIGURE 5 | BUB1-containing complexes. Combined Cytoscape web view of BUB1-containing complexes (squares) and their components (circles). The following terms were used to generate the display: complex terms (blue squares): BUB1:BUB3 (PR:000035566), BUB1:BUB3:MAD1L1 (PR:000035567), BUB1:BUB3:APC (PR:000035576), BUB1:BUB1B (PR:000035578), and BUB1:PLK1 (PR:000035580). Component terms (blue circles): hBUB1 (PR:000035400), hBUB3 (PR:000026899), yBUB1 (PR:000035402), yBUB3 (PR:000035532), SpomBUB1 (PR:000035570), SpomBUB3 (PR:000035571), hMAD1L1 (PR:000035474), hAPC (PR:000030190), hBUB1B (PR:000026903), hBUB1B/Phos:2 (PR:000035428), hBUB1B/PhosRes- (PR:000035373), hBUB1/Phos:7 (PR:000035412), and hPLK1 (PR:000035455). Species-specific BUB1:BUB3 complexes are shown in orange. Dotted arrows indicate the has_part relation; solid arrows indicate the is_a relation. Display options were set to show parents and all children, including organism level terms; nodes for siblings of complex components and complexes not containing BUB1 were hidden.

Ontology ID GO:0000776 (kinetochore). Although the kinetochore and centromere are distinct structures, the terms are sometimes used interchangeably in the literature; therefore, we also retrieved human centromere localized proteins by searching for human proteins (Taxon ID 9606) annotated with the GO term GO:0000779 (condensed chromosome, centromeric region). The searches returned 34 results, including 28 kinetochore-localized protein forms, 5 centromere localized forms, and one term – AURKB (PR:000035358) that is annotated with both kinetochore and centromere localization terms. These terms are annotated with PPI data mined from the literature and from several PPI databases. We downloaded the OBO stanzas and PAF for these proteins from PRO and used the information therein to build a network with Cytoscape (**Figure 6**). In addition to the PPIs (green arrows), the network displays kinases for the phosphorylated protein forms (blue arrows) and gene-level parent terms for the modification-level terms (black arrows).

Because functional annotation of PRO terms is an ongoing process, the set of kinetochore/centromere localized proteins we retrieved is not comprehensive nor is the PRO annotation of PPIs for these proteins complete. However, it is representative of the diverse functions of the kinetochore. The core checkpoint proteins BUB1, BUB1B, BUB3, MAD1L1, and MAD2L1 (**Figure 6**, red nodes) are found at the kinetochore/centromere and interact extensively with each other. All of the possible pair-wise interactions among these proteins are present except for BUB1B-MAD1 and BUB1-MAD2. The core checkpoint protein AURKB (purple) associates with this sub-network via an association with BUB1B. The checkpoint target CDC20 is also found at kinetochores/centromeres where it interacts with the MCC components MAD2, BUB1B, and BUB3. BUB1-dependent phosphorylation of CDC20 does not affect its ability to bind other MCC components as both CDC20/Phos:1 and CDC20/PhosRes-interact with MAD2 and BUB1B.

The checkpoint proteins are integrated into the larger environment of the kinetochore through interactions with other kinetochore/centromere proteins. AURKB binds to BIRC5, CDCA8, and INCENP (**Figure 6**, blue nodes) to form the Chromosomal

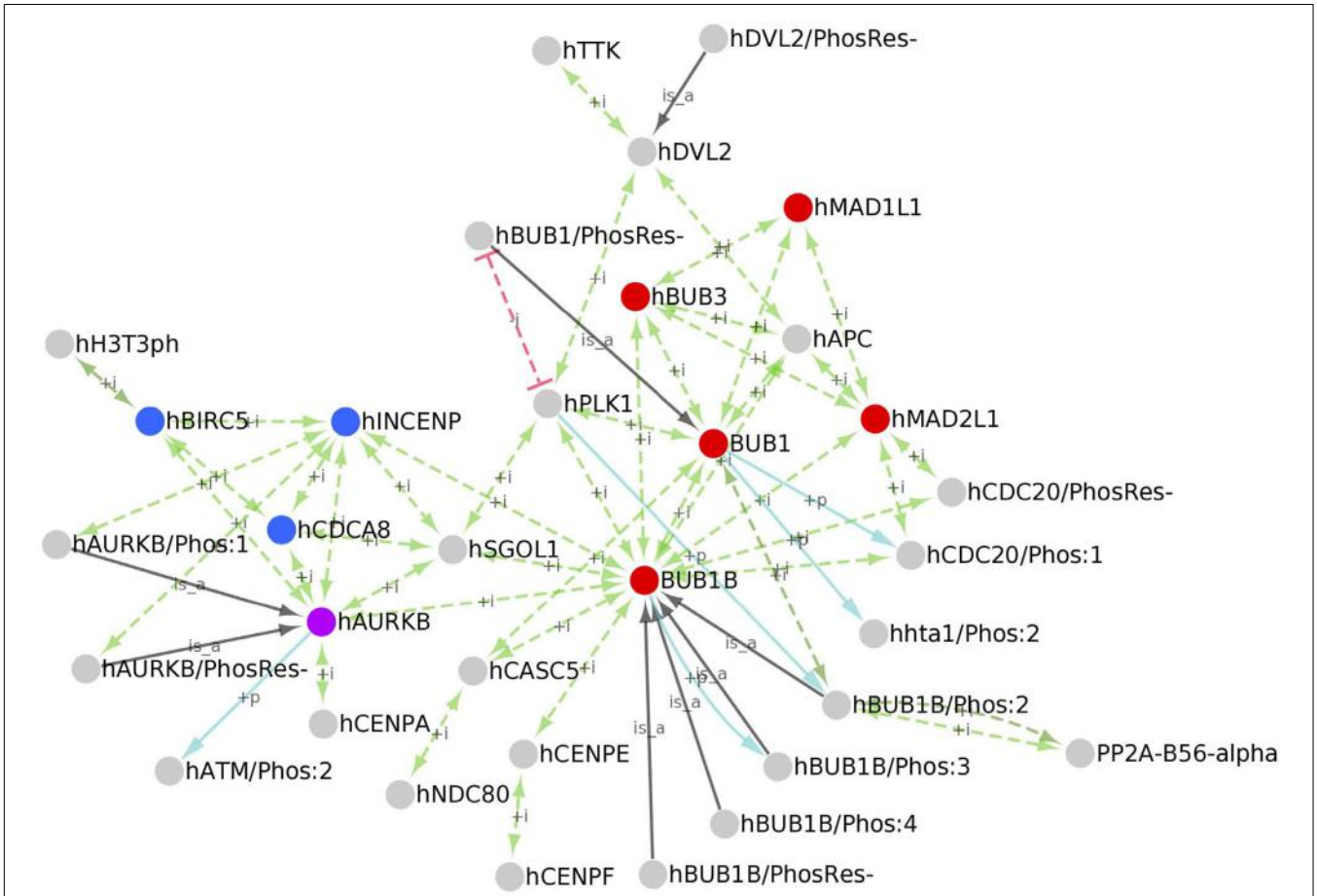


FIGURE 6 | PPI network of kinetochore/centromere localized proteins. Cytoscape network of the kinetochore/centromere localized proteins in PRO. PPIs (green edges), inhibited PPIs (red edges), kinase/phosphorylated product relationships (blue edges), and parent-child relationships for phosphorylated forms (black edges) are shown. Nodes representing the core spindle checkpoint proteins BUB1, BUB1B, BUB3, MAD1, and MAD2 are red; nodes representing the Chromosomal Passenger Complex (CPC) subunits INCENP, CDCA8, and BIRC5 are blue; AURKB is purple.

Passenger Complex (CPC; van der Waal et al., 2012). Both phosphorylated (AURKB/Phos:1) and unphosphorylated (AURKB/PhosRes-) forms interact with INCENP, suggesting that AURKB phosphorylation does not play a role in CPC formation. Several CPC subunits (AURKB, INCENP, and CDCA8) interact with SGOL1, a protein that participates in sister chromatid cohesion [see PRO annotation for SGOL1 (PR:000035551)]. The CPC is tethered to the centromere via interactions with centromeric histone subunits. In particular, the CPC subunit AURKB interacts with the centromeric histone H3 variant CENPA and BIRC5 interacts with the Thr-3 phosphorylated form of histone H3 (H3T3ph).

BUB1 and BUB1B both associate with the outer kinetochore component, CASC5. BUB1B makes other connections to the kinetochore via SGOL1 and CENPE, a protein that assists in the alignment of chromosomes on the metaphase plate [see PRO annotation for CENPE (PR:000035367)]. BUB1B binding to CENPE may stimulate its auto-phosphorylation activity [see BUB1B/Phos:3 (PR:000035432)].

Several spindle checkpoint proteins – BUB1, BUB1B, BUB3, and MAD2 – interact with APC and the checkpoint kinase TTK interacts with DVL2. APC and DVL2, which interact with each other, both participate in spindle assembly [see PRO annotation for APC (PR:000030190) and DVL2 (PR:000035487)]. The significance of these interactions is unclear, but it could reflect a role for APC and DVL2 in checkpoint signaling or a role for the checkpoint proteins in spindle assembly.

A protein kinase, PLK1, and a protein phosphatase, PP2A, associate with checkpoint proteins and other kinetochore proteins, positioning them to regulate critical kinetochore substrates. PLK1 interacts with the checkpoint proteins BUB1 and BUB1B as well as SGOL1 and DVL2. PLK1 association with BUB1 depends upon the prior phosphorylation of BUB1 (Qi et al., 2006); this dependence is represented in the network by a red line indicating an inhibited interaction between unphosphorylated BUB1 (BUB1/PhosRes-) and PLK1. As previously discussed, PLK1 phosphorylates BUB1B [see BUB1B/Phos:2 (PR:000035428)]. Additional kinetochore-localized substrates of PLK1 most likely exist. For example, human BUB1 is phosphorylated by CDK1 and PLK1 in a manner analogous to BUB1B/Phos:2 (BUB1/Phos:8; PR:000035418); however, further experiments are necessary to

show that this form is indeed localized to kinetochores and to determine its role in spindle assembly and checkpoint function. The phosphatase PP2A localizes to kinetochores by binding to phosphorylated BUB1B (BUB1B/Phos:2).

THE ROLE OF PROTEIN PHOSPHORYLATION AT THE KINETOCHORE/CENTROMERE

Eight of the kinetochore/centromere localized protein forms in our set are phosphorylated: BUB1B/Phos:2, BUB1B/Phos:3, BUB1B/Phos:4, AURKB/Phos:1, CDC20/Phos:1, ATM/Phos:2, H3T3ph, and HHTA1/Phos:2. Because phosphorylation can have a wide range of effects on proteins, affecting localization, function, and/or the processes in which they participate, we wanted to investigate the impact of phosphorylation on these particular proteins.

What functions, processes, and subcellular localizations are affected by protein phosphorylation in the human kinetochore?

In the PRO annotation, localizations, functions, and processes that are affected by protein modification are denoted by adding a modifier (such as increased or decreased) to the corresponding GO term and inclusion of a reference form. Thus, we searched for human proteins (Taxon ID 9606) localized to the kinetochore (Ontology ID GO:0000776) with at least one line of functional annotation that included a modifier (Modifer NOT NULL); to limit the results to phosphorylated proteins, we selected "Phosphorylated forms" from the Quick Links menu. For the reasons described above, we repeated the search substituting GO:0000779 (condensed chromosome, centromeric region) in the Ontology ID field. All eight of the kinetochore/centromere localized proteins appeared in our search results, indicating that all of these proteins had at least one attribute that was affected by phosphorylation. We examined the annotation for each protein and summarized the affected attributes in **Table 1**.

Even though phosphorylation is often used as a mechanism to regulate protein localization, none of the phosphorylated proteins in this group was annotated to indicate increased or decreased localization to the kinetochore/centromere relative the unphosphorylated form. In fact, the unphosphorylated forms of several of these proteins – BUB1B, CDC20, and AURKB – have been shown to localize to kinetochores with similar affinity as the phosphorylated forms see PRO annotation for BUB1B/PhosRes-(PR:000035373), CDC20/PhosRes-(PR:000035369), and AURKB/PhosRes-(PR:000 035661). Intriguingly, the kinases for CDC20/Phos:1 (kinase is BUB1), BUB1B/Phos:2 (kinase is PLK1), ATM/Phos:2 (kinase is AURKB), and HHTA1/Phos:2 (kinase is BUB1), are themselves kinetochore/centromere localized proteins (see **Figure 6**). In addition, BUB1B/Phos:3 phosphorylation depends on the association of BUB1B with the kinetochore-localized protein, CENPE. Taken together, these observations suggest that phosphorylation may occur after kinetochore localization. It would be interesting to test this hypothesis and to see if it holds true for a wider range of phosphorylated kinetochore/centromere localized proteins.

While phosphorylation did not affect the ability of these proteins to localize to the kinetochore/centromere themselves, three phosphorylated protein forms (phospho-Ser-121-Histone H2A, phospho-Thr-3-Histone H3, and BUB1B/Phos:3) showed an increased ability to recruit other proteins to the kinetochore/centromere relative to their respective unphosphorylated forms. Phosphorylation of Histone H2A on Ser-121 (HHTA1/Phos:2) creates a binding site for SGOL1. Phosphorylation of Histone H3 on Thr-3 (H3T3ph) creates a binding site for BIRC5, which in turn recruits the rest of the CPC (AUKB, CDCA8, and INCENP). Finally, BUB1B/Phos:3 is required for the kinetochore recruitment of MAD1 and MAD2.

Phosphorylation of BUB1B (BUB1B/Phos:2, BUB1B/Phos:3, and BUB1B/Phos:4) and AURKB (AURKB/Phos:1) is important for the ability of these proteins to regulate microtubule/kinetochore attachments as the phosphorylated forms show increased participation in attachment of spindle microtubules to kinetochores, metaphase plate congression, and/or chromosome segregation. Formation of stable, bipolar microtubule-kinetochore attachments requires a balance of kinase and phosphatase activity. AURKB destabilizes incorrect attachments by phosphorylating kinetochore components such as NDC80; the phosphatase PP2A counterbalances AURKB activity by dephosphorylating NDC80, thereby stabilizing attachments (Zich and Hardwick, 2010; Foley et al., 2011). Because AURKB kinase activity is important for destabilizing incorrect kinetochore-microtubule attachments, the increased kinase activity of AURKB/Phos:1 may explain its enhanced role in this process (Zich and Hardwick, 2010). On the other hand, BUB1B/Phos:2 may help stabilize nascent kinetochore-microtubule attachments through its increased affinity for PP2A, the phosphatase that reverses AURKB phosphorylation of NDC80 (Foley et al., 2011). Although its interaction with PP2A has not been directly assessed, BUB1B/Phos:3 shows a decreased ability to negatively regulate NDC80 phosphorylation (i.e., NDC80 phosphorylation is increased in the presence of BUB1B/Phos:3). This suggests that BUB1B/Phos:3 might have a reduced affinity for PP2A relative to unphosphorylated BUB1B. It would be interesting to test whether BUB1B/Phos:4 also affects the NDC80 phosphorylation/dephosphorylation cycle. Overall, these results suggest that BUB1B affinity for PP2A and consequently, the stability of kinetochore-microtubule attachments may be sensitively modulated by the BUB1B phosphorylation state.

Four proteins – Cdc20/Phos:1, AURKB/Phos:1, ATM/Phos:2, and BUB1B/Phos:3 – show an increased ability to mediate the spindle checkpoint relative to their unphosphorylated counterparts. CDC20/Phos:1 (phosphorylated by BUB1) shows decreased ubiquitin ligase activity relative to unphosphorylated CDC20, which presumably leads to its increased checkpoint activity. Thus, the spindle checkpoint acts through CDC20 in two independent ways to inhibit the APC/C: through formation of the MCC (BUB1B, BUB3, MAD2, and CDC20), which binds and inhibits the APC/C, and by phosphorylation of CDC20, which inhibits its ubiquitin ligase activity. Both AURKB/Phos:1 and ATM/Phos:2 have increased protein kinase activity relative to the unphosphorylated forms, which may be important in their increased ability to participate in the checkpoint response, although this possibility has not been directly tested. BUB1B/Phos:3 may participate in the checkpoint through its recruitment of MAD1L1 and MAD2L1 to kinetochores.

DISCUSSION

The structural framework and features of PRO enable the investigation of many aspects of proteins and complexes, particularly analyses of cross-species relationships and relationships between modified proteins forms and functions. Our spindle checkpoint use case outlines a number of strategies that can be generalized to other cellular processes or pathways of interest.

INVESTIGATION OF THE ROLE OF MODIFIED PROTEIN FORMS IN A BIOLOGICAL PROCESS

In this study we showed how the PRO framework could be used to investigate the role of different protein forms that participate in a biological process of interest. We focused on PTM protein forms, as PTM is a central mechanism for the regulation of protein function in cells. Most PTM resources specialize in a single type of modification (e.g., phosphorylation) and are organized around individual modification sites. However, protein modification *in vivo* is usually a combinatorial process where proteins are subject to multiple types of modifications on multiple sites. In this regard, PRO offers a more realistic view of protein modification through its representation of protein forms that carry the combinations of modifications that are observed *in vivo*. The representation of protein complexes in PRO also takes into account the modification state of the complex components. Moreover, modified forms and complexes in PRO can be individually annotated with functional information, making it possible to discern the contribution of each to a biological process.

We used PRO to explore the role of protein phosphorylation in the context of the spindle checkpoint. Our examination of the PRO representation of human BUB1B phosphorylated forms and complexes revealed multiple phosphorylated forms of this protein and at least two participating kinases (**Figure 3**). Comparison of the annotation of the BUB1B phosphorylated forms provided an additional level of information that revealed some intriguing phosphorylation state-dependent differences in function. For example, BUB1B/Phos:2 and BUB1B/Phos:3 have opposite effects on the phosphorylation of the kinetochore protein, NDC80 (**Table 1**).

Table 1 | Functional effects of phosphorylation of kinetochore/centromere localized proteins.

Protein	Modifier	Function/process	Targets
CDC20/Phos:1	Decreased	Ubiquitin protein ligase activity	
	Increased	Spindle checkpoint	
BUB1B/Phos:2	Increased	Protein binding	BUB1
	Increased	Protein binding	PP2A
	Increased	Protein kinase activity	
	Increased	Attachment of spindle microtubules to kinetochores	
	Increased	Metaphase plate congression	
BUB1B/Phos:3	Increased	Metaphase plate congression	
	Increased	Chromosome segregation	
	Increased	Spindle checkpoint	
	Increased	Protein localization to kinetochore	MAD1L1, MAD2L1
	Decreased	Negative regulation of protein phosphorylation	NDC80
BUB1B/Phos:4	Increased	Attachment of spindle microtubules to kinetochores	
	Increased	Inhibition of mitotic anaphase-promoting complex activity	
	Increased	Metaphase plate congression	
AURKB/Phos:1	Increased	Protein kinase activity	
	Increased	Chromosome segregation	
	Increased	Metaphase plate congression	
	Increased	Spindle checkpoint	
ATM/Phos:2	Increased	Protein kinase activity	
	Increased	Spindle Checkpoint	
HHTA1/Phos:2	Increased	Protein localization to chromosome, centromeric region	SGOL1
H3T3ph	Increased	Protein binding	BIRC5
	Increased	Protein localization to chromosome, centromeric region	AURKB, CDCA8, INCENP, BIRC5

PRO terms retrieved using the search query: "Taxon ID 9606 (human) AND Ontology ID GO:0000776 (kinetochore) AND Modifier NOT NULL OR Taxon ID 9606 (human) and Ontology ID GO:0000779 (condensed chromosome, centromeric region) and Modifier NOT NULL." Results were restricted to phosphorylated proteins only by selecting "Phosphorylated forms" from the Quick Links menu. Protein binding partners were obtained from the "Interaction with" column of the Functional Annotation section of the PRO entry page. Targets for the annotation terms "Protein localization to the kinetochore," "Protein localization to the chromosome, centromeric region," and "Negative regulation of protein phosphorylation" were obtained from the comment column of the PAF.

Moreover, in an analysis of phosphorylated protein forms that localize to the kinetochore, we found that phosphorylation did not enhance or suppress kinetochore localization *per se*, but did affect the ability of proteins to recruit other proteins to the kinetochore (**Table 1**). Finally, we found that multiple BUB1B forms form complexes with BUB1 (**Figure 5**).

CROSS-SPECIES COMPARISON OF MODIFIED PROTEIN FORMS

A related biological question that can be addressed with PRO concerns the cross-species conservation of modified protein forms. Here we described a small scale study involving the phosphorylation of one protein – BUB1B – in three organisms – human, frog, and mouse. Based on the descriptions of BUB1B phosphorylated forms in PRO and a multiple sequence alignment, we concluded that all four BUB1B phosphorylated forms found in humans could be conserved in mice. Three of the four forms are either known to be conserved in frogs or are likely to be, but one form, BUB1B/Phos:4, is not.

Discovery that a modified form found in one species is not conserved in another species is very interesting because a comparison of the function of that protein in the two organisms can provide insight into the role of the modification. Prediction that a modified protein form is conserved in a species where it has not yet been characterized is also useful because it expands the pool of organisms that can be used to study the modified form. For example, confirmation of the existence of BUB1B phosphorylated forms in mice would allow the study of BUB1B forms in mammalian cells undergoing meiosis. These studies could shed light on a question about the function of BUB1B/Phos:1. Frog BUB1B/Phos:1 has been shown to be required for spindle checkpoint cell cycle arrest; in contrast, human BUB1B/Phos:1 is dispensable for cell cycle arrest under these circumstances (Elowe et al., 2007; Wong and Fang, 2007). It is unclear whether this indicates a true difference between the human and frog BUB1B proteins, or if it reflects the fact that the human checkpoint was tested in mitotically growing cells, whereas the frog checkpoint was tested in extracts of oocytes undergoing meiosis. If BUB1B/Phos:1 is indeed present in mice, it would be very interesting to compare its involvement in checkpoint arrest during mitosis and meiosis.

Cross-species analysis of modified protein forms is not limited to a single protein. It can be expanded to include all modified proteins involved in a biological process or present in a particular cellular compartment. It is also not restricted to phosphorylated proteins. The PRO framework can be used to define many kinds of modified protein forms, including those that arise from post-translational modifications such as methylation, acetylation, and ubiquitination and protein isoforms that arise from alternative splicing or from protein cleavage.

ANALYSIS OF EVOLUTIONARY RELATIONSHIPS AMONG PROTEINS

Research into the mechanisms of a biological process often proceeds simultaneously in multiple model systems. In many cases, a clear picture of the process emerges only after data generated from disparate lines of experiment are considered as a whole. Merging of data in this way relies on the assumption that the proteins and pathways examined in the different systems are functionally related. The organization of PRO reflects evolutionary relationships among proteins and can be used as a guide in cross-species comparisons of experimental results. In PRO, organism-specific terms that share 1:1 orthology are grouped under a species-independent parent term (gene-level term) and species-independent terms that share a common domain structure are further grouped under a family level terms. In our analysis, we found that human and yeast BUB1 are 1:1 orthologs and thus share the same species-independent parent terms. However, human BUB1B lies on a separate branch of the PRO hierarchy from its closest yeast relative, MAD3 (**Figure 2**). Thus, assumptions about the conservation of BUB1 function in humans and yeast are more easily justified than assumptions about the conservation of BUB1B/MAD3 function. Similarly, PRO complexes are grouped under a species-independent complex term if their components are orthologous. Our examination of the BUB1/BUB3 complex revealed that it is conserved in budding yeast, fission yeast, and humans (**Figure 5**).

CONSTRUCTION OF PPI NETWORKS

Often, it is possible to gain insight into the function of proteins in a common pathway by examining their PPIs. PRO facilitates the construction of PPI networks for groups of proteins that are related by some common attribute. Using the built-in PRO search function, it is possible retrieve all PRO terms that share an attribute (e.g., kinetochore localization). The PAF for these terms, which contains PPI information in machine-readable format, can then be downloaded and used to build a PPI network with Cytoscape. Because PRO annotation can show interactions that are dependent on protein modification, PPI networks constructed with PRO have an added dimension that is absent from other PPI network building resources. For example, our PPI network of kinetochore-localized proteins shows that PP2A-B56-alpha interacts specifically with BUB1B/Phos:2 and that PLK1 fails to interact with the unphosphorylated form of BUB1 (**Figure 6**).

CONCLUSION

As we have shown with this use case, PRO is a valuable tool for the study of a complex biological process. Interoperating with other ontologies and resources, PRO provides a structural framework that organizes current knowledge about protein forms, complexes, and cross-species relationships among proteins. While we focused on the spindle checkpoint, the PRO search, display, and analysis strategies we demonstrated here can be applied to any process. PRO-based analysis is particularly valuable for processes where modified protein forms play a prominent role. While PRO coverage is limited for modified forms, we rely on the user community to help in populating the ontology. The web-based RACE-PRO interface provides one means for the user to contribute to PRO. As PRO grows, it will become an increasingly useful resource that can provide insight into biological processes and stimulate the generation of experimentally testable hypotheses.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation [ABI-1062520] and the National Institutes of Health [2R01GM080646-06].

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Bult, C. J., Drabkin, H. J., Evsikov, A., Natale, D., Arighi, C., Roberts, N., et al. (2011). The representation of protein complexes in the Protein Ontology (PRO). *BMC Bioinformatics* 12:371. doi:10.1186/1471-2105-12-371

Ceusters, W., and Smith, B. (2010). A unified framework for biomedical terminologies and ontologies. *Stud. Health Technol. Inform.* 160, 1050–1054.

Chatr-Aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., et al. (2007). MINT: the molecular INTeraction database. *Nucleic Acids Res.* 35, D572–D574.

Chen, R. H. (2002). BubR1 is essential for kinetochore localization of other spindle checkpoint proteins and its phosphorylation requires Mad1. *J. Cell Biol.* 158, 487–496.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39, D691–697.

D'Arcy, S., Davies, O. R., Blundell, T. L., and Bolanos-Garcia, V. M. (2010). Defining the molecular basis of BubR1 kinetochore interactions and APC/C-CDC20 inhibition. *J. Biol. Chem.* 285, 14764–14776.

Elowe, S., Dulla, K., Uldschmid, A., Li, X., Dou, Z., and Nigg, E. A. (2010). Uncoupling of the spindle-checkpoint and chromosome-congression functions of BubR1. *J. Cell. Sci.* 123, 84–94.

Elowe, S., Hummer, S., Uldschmid, A., Li, X., and Nigg, E. A. (2007). Tension-sensitive Plk1 phosphorylation on BubR1 regulates the stability of kinetochore microtubule interactions. *Genes Dev.* 21, 2205–2219.

Foley, E. A., Maldonado, M., and Kapoor, T. M. (2011). Formation of stable attachments between kinetochores and microtubules depends on the B56-PP2A phosphatase. *Nat. Cell Biol.* 13, 1265–1271.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2012). The units ontology: a tool for integrating units of measurement in science. *Database* 2012, bas033.

Goujon, M., Mcwilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., et al. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* 38, W695–W699.

Guo, Y., Kim, C., Ahmad, S., Zhang, J., and Mao, Y. (2012). CENP-E – dependent BubR1 autophosphorylation enhances chromosome alignment and the mitotic checkpoint. *J. Cell Biol.* 198, 205–217.

Hardwick, K. G., Johnston, R. C., Smith, D. L., and Murray, A. W. (2000). MAD3 encodes a novel component of the spindle checkpoint which interacts with Bub3p, Cdc20p, and Mad2p. *J. Cell Biol.* 148, 871–882.

Hegemann, B., Hutchins, J. R., Hudecz, O., Novatchkova, M., Rameseder, J., Sykora, M. M., et al. (2011). Systematic phosphorylation analysis of human mitotic protein complexes. *Sci. Signal.* 4, rs12.

Hoehndorf, R., Dumontier, M., Oellrich, A., Rebholz-Schuhmann, D., Schofield, P. N., and Gkoutos, G. V. (2011). Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS ONE* 6:e22006. doi:10.1371/journal.pone.0022006

Hori, T., and Fukagawa, T. (2012). Establishment of the vertebrate kinetochores. *Chromosome Res.* 20, 547–561.

Huang, H., Hittle, J., Zappacosta, F., Annan, R. S., Hershko, A., and Yen, T. J. (2008). Phosphorylation sites in BubR1 that regulate kinetochore attachment, tension, and mitotic exit. *J. Cell Biol.* 183, 667–680.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846.

Lara-Gonzalez, P., Westhorpe, F. G., and Taylor, S. S. (2012). The spindle assembly checkpoint. *Curr. Biol.* 22, R966–R980.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics* 26, 2347–2348.

Matsumura, S., Toyoshima, F., and Nishida, E. (2007). Polo-like kinase 1 facilitates chromosome alignment during prometaphase through BubR1. *J. Biol. Chem.* 282, 15217–15227.

Montecchi-Palazzi, L., Beavis, R., Binz, P. A., Chalkley, R. J., Cottrell, J., Creasy, D., et al. (2008). The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.* 26, 864–866.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., et al. (2012). The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* 19, 190–195.

Natale, D. A., Arighi, C. N., Barker, W. C., Blake, J. A., Bult, C. J., Caudy, M., et al. (2011). The protein ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.* 39, D539–545.

Oh, H. J., Kim, M. J., Song, S. J., Kim, T., Lee, D., Kwon, S. H., et al. (2010). MST1 limits the kinase activity of aurora B to promote stable kinetochore-microtubule attachment. *Curr. Biol.* 20, 416–422.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–301.

Qi, W., Tang, Z., and Yu, H. (2006). Phosphorylation- and polo-box-dependent binding of Plk1 to Bub1 is required for the kinetochore localization of Plk1. *Mol. Biol. Cell* 17, 3705–3716.

Seeley, T. W., Wang, L., and Zhen, J. Y. (1999). Phosphorylation of human MAD1 by the BUB1 kinase in vitro. *Biochem. Biophys. Res. Commun.* 257, 589–595.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., et al. (2005). Relations in biomedical ontologies. *Genome Biol.* 6, R46.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432.

Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., et al. (2011). The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* 39, D698–704.

Suijkerbuijk, S. J., Van Dam, T. J., Karagoz, G. E., Von Castelmur, E., Hubner, N. C., Duarte, A. M., et al. (2012). The vertebrate mitotic checkpoint protein BUBR1 is an unusual pseudokinase. *Dev. Cell* 22, 1321–1329.

Sun, S. C., and Kim, N. H. (2012). Spindle assembly checkpoint and its regulators in meiosis. *Hum. Reprod. Update* 18, 60–72.

Taylor, S. S., Hussein, D., Wang, Y., Elderkin, S., and Morrow, C. J. (2001). Kinetochore localisation and phosphorylation of the mitotic checkpoint components Bub1 and BubR1 are differentially regulated by spindle events in human cells. *J. Cell. Sci.* 114, 4385–4395.

van der Waal, M. S., Hengeveld, R. C., Van Der Horst, A., and Lens, S. M. (2012). Cell division control by the chromosomal passenger complex. *Exp. Cell Res.* 318, 1407–1420.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191.

Wong, O. K., and Fang, G. (2007). Cdk1 phosphorylation of BubR1 controls spindle checkpoint arrest and Plk1-mediated formation of the 3F3/2 epitope. *J. Cell Biol.* 179, 611–617.

Yuan, X., Hu, Z. Z., Wu, H. T., Torii, M., Narayanaswamy, M., Ravikumar, K. E., et al. (2006). An online literature mining tool for protein phosphorylation. *Bioinformatics* 22, 1668–1669.

Zich, J., and Hardwick, K. G. (2010). Getting down to the phosphorylated 'nuts and bolts' of spindle checkpoint signalling. *Trends Biochem. Sci.* 35, 18–27.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 February 2013; accepted: 05 April 2013; published online: 26 April 2013.

Citation: Ross KE, Arighi CN, Ren J, Natale DA, Huang H and Wu CH (2013) Use of the protein ontology for multi-faceted analysis of biological processes: a case study of the spindle checkpoint. Front. Genet. ***4****:62. doi: 10.3389/fgene.2013.00062*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 11 October 2013
doi: 10.3389/fgene.2013.00201

Annotation extension through protein family annotation coherence metrics

Hugo P. Bastos[1]*, Luka A. Clarke[2] and Francisco M. Couto[1]

[1] LaSIGE, Department of Informatics, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[2] Department of Chemistry and Biochemistry, BioFIG - Centre for Biodiversity, Functional and Integrative Genomics, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

Edited by:
John Hancock, University of Cambridge, UK

Reviewed by:
Zoran Nikoloski, Max-Planck Institute of Molecular Plant Physiology, Germany
Pascale Gaudet, Swiss Institute of Bioinformatics, Switzerland

***Correspondence:**
Hugo P. Bastos, Department of Informatics, Faculdade de Ciências, Universidade de Lisboa, Bloco C6, Sala 6.3.33, Campo Grande, Lisboa 1749-016, Portugal
e-mail: hbastos@xldb.di.fc.ul.pt

Protein functional annotation consists in associating proteins with textual descriptors elucidating their biological roles. The bulk of annotation is done via automated procedures that ultimately rely on annotation transfer. Despite a large number of existing protein annotation procedures the ever growing protein space is never completely annotated. One of the facets of annotation incompleteness derives from annotation uncertainty. Often when protein function cannot be predicted with enough specificity it is instead conservatively annotated with more generic terms. In a scenario of protein families or functionally related (or even dissimilar) sets this leads to a more difficult task of using annotations to compare the extent of functional relatedness among all family or set members. However, we postulate that identifying sub-sets of functionally coherent proteins annotated at a very specific level, can help the annotation extension of other incompletely annotated proteins within the same family or functionally related set. As an example we analyse the status of annotation of a set of CAZy families belonging to the Polysaccharide Lyase class. We show that through the use of visualization methods and semantic similarity based metrics it is possible to identify families and respective annotation terms within them that are suitable for possible annotation extension. Based on our analysis we then propose a semi-automatic methodology leading to the extension of single annotation terms within these partially annotated protein sets or families.

Keywords: functional annotation, annotation extension, protein annotation coherence, annotation metrics, gene ontology

1. INTRODUCTION

The continuous development of high-throughput methodologies for biological molecule sequencing has led to an increase in the amount of raw biological data in need of further processing. The sequencing of a new biological molecule is normally followed by a functional annotation process that aims to provide functional descriptors elucidating its biological role. Functional annotations can be derived from either experimental determination or prediction. Generically, given supporting evidence, functional descriptors are assigned (with varying degrees of confidence) to their corresponding biomolecules. In fact, a functional annotation can be represented as the pair of a biomolecule (identifier) and corresponding functional descriptor.

Among biomolecules, proteins are of particular interest given their participation in practically every process occurring within living cells. Their functions can range from structural or mechanical support to the catalysis of vital metabolic biochemical reactions. Furthermore, their functional specification is very broad and can range from descriptors on general participation in biological processes, such as responses to oxidative stress, up to more specific descriptors, such as catalysis of particular biochemical reactions. It would be desirable to determine protein function via accurate and comprehensive chemical characterizations, if possible by experimental assessment, however, this process is expensive and time consuming. Instead, the most commonplace approach is the use of any of the several function prediction methodologies, relying on techniques ranging from sequence homology detection to text mining of the scientific literature. Most of these methodologies also rely heavily on computational power and can range from partial to full automation, thus enabling them to handle the barrage of biological sequence data currently being made available.

Proteins are commonly grouped into evolutionarily related groups known as *protein families*. Within a family each protein shares homology with all the other proteins, i.e., it descends from a common ancestor and usually retains significant sequence similarity. In turn that often (but not always) translates into similar three-dimensional structures and functions. Although sequence similarity alone is not sufficient to conclude protein homology, it nevertheless provides a reasonable cornerstone for many sequence alignment methods. Similarly, homology also does not guarantee functional similarity among proteins but provides a good starting point and is commonly used in several functional annotation methods. Hence, it is typically advantageous to group proteins into homologous families because of the potentially shared functions.

The emergence of biological ontologies and most notably the Gene Ontology (GO) (Ashburner et al., 2000) has greatly benefited the annotation efforts by providing a structured and controlled vocabulary of terms for the description of gene products.

This standardization of human-readable functional descriptors also enables machine-readability thus being particularly useful in automated procedures. This in turn leads to an ever increasing availability and quality of protein annotations. The increasing popularity of GO terms for protein annotation has also led to the development of several associated semantic similarity based metrics that compare proteins based on their functions instead of their sequence or structure. GO semantic similarity can then be defined as the closeness in meaning between two terms or two sets of terms annotating two proteins. Under the assumption that when functional descriptors of two proteins are similar so are their functions, semantic similarity is then also referred to as *functional similarity*. However, caution must be exerted during the interpretation of annotation similarities since there are still issues that GO inherently does not solve, for instance, annotation bias and annotation incompleteness.

The functional descriptions of GO, given its ongoing and asymmetric growth, span a range of specificities (Alterovitz et al., 2010). Coupled with that, protein prediction methods assign either more specific or more generic annotation terms depending on the uncertainty level of the predictions being made. When comparing proteins annotated at different levels of completeness low semantic similarity values may then be reported. Therefore, the metrics used either have to account for these issues or adequate care must be taken when interpreting results.

The development of functional similarity metrics able to explicitly gauge the state of annotation incompleteness within a set of functionally related proteins is much required. We further postulate that by implementing these kind of metrics, we can identify functionally coherent sub-sets of proteins with a greater degree of annotation "completeness." Using these identified subsets as specific function knowledgebases we can potentiate the annotation extension of the remaining members in a functionally related set that is still incompletely annotated, ultimately leading to a greater degree of annotation completeness for a given functionally related protein set.

2. THEORY

2.1. GENE ONTOLOGY

The GO consortium provides a structured and controlled vocabulary for the description of molecular phenomena in which proteins (and or gene products) are involved. Within each GO aspect the biological phenomena are described at different levels, thus this vocabulary is divided into three orthogonal ontology aspects that describe gene products in terms of their associated biological processes, cellular components and molecular functions (Ashburner et al., 2000). The *biological process* aspect of GO describes activities of sets of proteins interacting and involved in cellular processes, such as metabolism or signal transduction. The cellular localizations (such as the Golgi complex or the ribosome), where these processes take place are described by the *cellular component* aspect of the ontology. On the other hand, each protein can, independent of the surrounding environment, perform catalytic or binding elementary molecular activities thus being described by the *molecular function* aspect of the ontology. Structurally each ontology aspect is organized as a Directed Acyclic Graph (DAG), where each node represents a term and edges represent a relationship between those terms. Each term is identified by an alphanumeric code (e.g., GO:0001170) and its textual descriptors, including its name, definition, and synonyms if available. Currently, the existing relationships between GO terms can be of three types: *is_a*, *part_of* and *regulates*. While *is_a* and *part_of* relations are only established within each individual ontology aspect, *regulates* relations can occur across aspects.

Proteins and other gene products are not actually part of GO which includes only terms that describe them. Nevertheless, the GO Consortium, via the Gene Ontology Annotation (GOA) project (Barrell et al., 2009), does provide annotations, such as previously defined as being the associations between gene products and the GO terms that functionally describe them. In order to fully describe a protein function any number of GO terms can be used to annotate the protein. Additionally, GO follows the true path rule which states that "the pathway from a child term all the way up to its top-level parent(s) must always be true", thus as can be seen in **Figure 1** any protein annotated to the term *polysaccharide binding* is also automatically annotated to its two parent terms: *carbohydrate binding* and *pattern binding*. In turn these two sibling terms are children of the term *binding*, a direct child of the root term *molecular_function*. Furthermore, each annotation linking a GO term to a protein is given an evidence code (ECO), which is an acronym identifying the type of evidence that supports that annotation, e.g., the IDA code (Inferred by Direct Assay) is assigned to annotations that are supported by that type of experiment.

2.2. PROTEIN GO ANNOTATION

Functional annotation is an essential step in the path of providing proteins their biological contexts and therefore facilitates

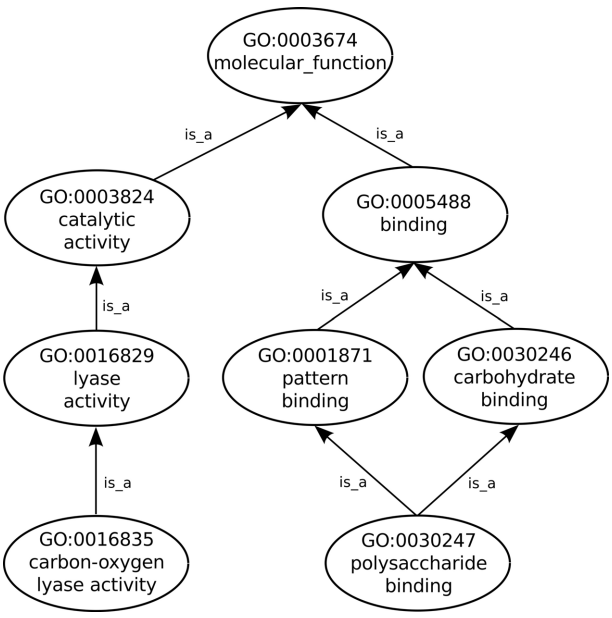


FIGURE 1 | Sub-graph of the GO *molecular function* ontology aspect depicting only *is_a* relationships.

knowledge exchange within the scientific community. Several methodologies exist for protein annotation but generally they can be divided into three major approaches: manual annotation (or curation), automatic annotation and the hybrid approach, semi-automatic annotation. Despite manual annotations produced by expert curators typically being of an higher quality level, this annotation approach does not scale up to the output of the high-throughput sequencing projects. Therefore, the bulk of protein annotations are produced via automated procedures. These typically rely on methods for transferring annotation terms from previously annotated protein sources to other unannotated (or incompletely annotated) proteins.

Using a controlled vocabulary like GO for protein annotation instead of free-text annotation solves several issues mostly common to many early annotation systems. Among those issues are the lack of annotation interoperability due to researcher subjectivity, the lack of vocabulary uniformity and problems arising from different scopes in function definitions. The scope of annotation can range from gene identification, cellular component specification and description of molecular interactions up to regulatory interactions between components of whole biological systems. This could present itself as an issue during the annotation process but is dealt with by the GO structure, where these scopes are divided into three orthogonal ontology aspects: *cellular component*, *molecular function*, and *biological process*. However, GO does not solve all annotation issues and even introduces new ones. The GO ontology aspects themselves are a product of mostly manual curation and their growth is linked to research bias, thus some parts of the ontology are more developed (have terms for more specific functions) than others (Pesquita and Couto, 2012). This is a source of incompleteness for annotation by limiting the maximum functional specificity that can be attributed to proteins. However, despite the availability of specific terms some annotation methodologies (mostly automatic) are unable to use them to annotate proteins with an high degree of confidence. Hence, this leads to a similar type of annotation incompleteness. On the other hand, a conservative annotation behavior may be desirable in order to mitigate possible annotation error propagation.

Typically, the automatic protein annotation systems do not actually produce *de novo* functional annotation terms. Instead, these systems commonly rely on methods for transferring annotation terms from previously annotated protein sources to other unannotated (or incompletely annotated) proteins. Thus, the typical workflow of an automatic annotation system includes a first stage where potential functional peers are identified. A second stage then involves the actual annotation transfer where functional terms are extracted from the functional peers and associated to the previously incompletely annotated or unannotated proteins.

The automatic procedures used for protein annotation can be divided into sequence-based approaches and structure-based approaches. Although three dimensional structure of proteins is generally more conserved than its sequence, the wider availability of sequence data over structural data allows for potentially greater annotation coverage with the former. Still, proteins with similar sequences typically possess evolutionary proximity, and to some extent, function conservation thus providing good approximations. In a similar sense, structure-based approaches can also compare protein structures in order to obtain similarity scores, but further details on structure-based approaches are out of the scope of this topic (see more at Sleater and Walsh, 2010).

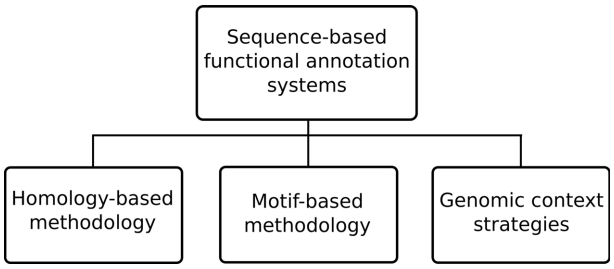


FIGURE 2 | Common approaches in sequence-based functional annotation systems.

The sequence-based approaches can still be further subdivided, as depicted in **Figure 2**, into three specific methodology types: homology-based, motif-based and genomic context strategies. Among the existing functional annotation systems the homology-based methodology is perhaps the most prevalent methodology. This type of methodology generally makes use of sequence alignment algorithms, such as the ubiquitous BLAST (Altschul et al., 1997), to compare unannotated query proteins against annotated sequences in a database. The underlying assumption is that similar sequences are most likely to have evolved from a common ancestor and thus retained similar functions. However, high sequence similarity does not always mean functional similarity (Rost, 2002) so annotation systems also employ additional techniques to handle known caveats. An example of a system using this approach is Blast2GO (Götz et al., 2008) where homologous sequences are retrieved from significant BLAST results under a given expectation value (*e*-value) threshold. In order to handle the possibility of annotation of short sequence matches with low *e*-values filtering by minimal alignment length (hsp-length) is allowed. An alternative to querying unannotated sequences against databases of annotated sequences, is to query them instead against known recurring patterns of motifs known to be associated with particular functions. This is the so-called sequence motif-based methodology where an annotation system uses either the patterns, rules and profiles of PROSITE (Sigrist et al., 2010), the fingerprints in PRINTS (Attwood, 2003), the family profiles from ProDom (Bru et al., 2005), the Hidden Markov Models (HMMs) from Pfam databases (Finn et al., 2010) or any other sequence motif type in order to perform functional inference.

Other alternative annotation strategies can be categorized under the denomination of genomic context strategies. These strategies subsume the gene neighborhood, gene clustering, Rosetta stone and phylogenetic profiles methods, which operate by identifying pairs of non-homologous proteins that co-evolve. Evolutionary pressure originates pairs of proteins that functionally collaborate and that: (i) are coded nearby in multiple genomes, (the gene neighborhood method); (ii) are components

of an operon in prokaryotes, (the gene cluster method); (iii) can be fused into a single protein in some organisms, (the Rosetta stone method); (iv) are regularly both present or both absent within genomes, (the phylogenetic profiles method) (Bowers et al., 2004). Protein-protein interactions and gene expression data from microarray experiments have also been used as part of the functional peers identification methodology in some annotation systems. These genomic context methods can be used on annotation systems either individually or conjointly. Overbeek et al. (1999) apply the gene clustering method on their system to infer functional coupling in prokaryotic genomes. Zheng et al. (2002) also uses a clustering method but applied on phylogenetic profiles. Using microarray mouse expression data for nearly 40,000 known and predicted mRNAs in 55 mouse tissues Zhang et al. (2004) were able to show that quantitative transcriptional co-expression is a powerful predictor of gene function. On the other hand, the Prolinks (Bowers et al., 2004) database uses the four genomic methods described above in combination to infer functional linkage between proteins through the identification of co-evolved pairs of non-homologous proteins. Similarly Phydbac (Enault et al., 2005), a gene function predictor system specialized in bacterial genomes also uses genomic context strategies in its workflow. Protein associations are generated by a combination of the phylogenetic profiles, the gene cluster and Rosetta stone methods. Both Deng et al. (2002) and Letovsky and Kasif (2003) employ the theory of Markov random fields to infer a protein's functions using protein-protein interaction data and the functional annotations of a protein's interaction partners. Chua et al. (2006) also developed a method for predicting protein function based on protein-protein interaction data, the difference being that in this case transitive relations are also considered for the predictions.

Prediction and assignment of protein function is seldom done in a deterministic way. While some general functions can be assigned deterministically to sequences, as protein function specificity rises the uncertainty of predicting an exact assignment does also. Thus, following the identification of functional peers it is common for annotation systems to employ an additional stage where term selection and transfer occurs. A confidence measure is usually associated with these term transfers, which often derives directly from probabilistic features from either statistical or machine learning methods employed for term selection, or alternatively, arbitrary empirical confidence measures from rule-based term selection methods. The methodologies used at the annotation transfer stage can be roughly grouped into three types: rule-based transfer, statistical transfer and machine learning transfer. One example of a rule-based methodology for annotation transfer occurs in the previously mentioned Blast2GO annotation system. There, for each candidate GO term, the highest similarity weighted by their ECO is considered. In addition the level of abstraction is also considered through the use of a rule counting the total number of GO terms unified at a given node weighted by a user set factor that controls the possibility and strength of abstraction. In the end, the annotation rule will only transfer the lowest terms in each branch that surpass an user defined threshold (Götz et al., 2008). A statistical-based annotation transfer methodology is used for example on the GOtcha (Martin et al., 2004) annotation system. GOtcha calculates probabilities for each term and its set of ancestors which allows some functions for a given sequence to be assigned with more confidence than others. Those probabilities are derived from two scores based on the expectation scores of pairwise matches between query sequences and database sequences and also the annotation distribution within each aspect of the GO ontology. On the other hand, the GOPET (Vinayagam et al., 2006) annotation system uses yet another type of approach: machine learning. In this system, GO terms associated to the retrieved homolog sequences are used in conjunction with several elaborate attributes, including sequence similarity measures, such as *e*-value, bit-score, identity, coverage score, and alignment length. Further attributes use GO-term frequency, GO term relationships between homolog sequences, the level of annotation within the GO hierarchy and homolog annotation quality which is calculated based on the ECO provided by the gene association tables of the GO mapped sequence databases. These attributes are used as training instances for support vector machines (SVM) which are then used to assign GO term annotation to the previously unannotated sequences.

2.3. SEMANTIC SIMILARITY

In the context of ontology, semantic similarity can be defined as the closeness in meaning between two ontology terms or two sets of terms annotating two entities represented by a given metric. Typically, the semantic similarity between two proteins annotated with GO terms is also called functional similarity, since it presents a measure of how similar the protein functions are.

Semantic similarity measures for comparing terms in an ontology typically rely on two main approaches: edge-based and node-based. Edge-based approaches in their most simple form rely on counting the number of edges between two terms on the ontology graph, which conveys a distance measure that can easily be converted to a similarity measure (Rada et al., 1989). Thus, the shorter the distance between two terms, the more similar they are. Different edges can have different associated semantic values leading to more sophisticated metrics. On the other hand node-based approaches can be better suited for ontologies such as GO, where nodes and edges are not uniformly distributed. A commonly used node property is the information content (IC), which is a frequency-based measure of how specific a term is within a given corpus (Resnik, 1995). Conveniently, the GOA project provides a suitable body of GO annotations that can used as a corpus. The *IC* of term can then be given by Equation 1.

$$IC(t) = -\log_2 f(t) \tag{1}$$

In Equation 1 $f(t)$ is the probability of annotation of term t. Consequently, terms annotating many proteins will score a low *IC*, while specific terms annotating only a few proteins will score an high *IC*. Additionally, the *IC* values can be normalized in order to provide a more intuitive meaning.

GO-based semantic similarity for proteins is given by the comparison of the sets of GO terms annotating each protein being compared within each GO ontology aspect. Two main approaches, pairwise and groupwise (Pesquita et al., 2009) are

typically used for this purpose. Pairwise approaches use semantic similarities between the GO terms annotating each protein, the semantic similarities are calculated for all possible pairs of terms between each set. Common among these approaches are variations such as the all pairs technique, where every pairwise combination is considered or the best pairs technique where only the best-matching pair for each term is considered. Global functional similarity scores between the actual proteins are usually obtained by averaging, summing or selecting the maximum of the pairwise similarity scores. For more on ontology-based semantic similarity check reviews by Pesquita et al. (2009) and Gan et al. (2013).

Several assessment studies have employed the developed semantic similarity measures for GO terms. There is no best measure for comparing terms, proteins or other gene products, it always depends on which specific task they are being used for. Lord et al. (2003) were among the first to assess the performance of different semantic similarity measures in the context of GO. For that purpose they adapted and tested three measures: Resnik's (Resnik, 1995), Lin's (Lin, 1998), and Jiang and Conrath's (Jiang and Conrath, 1997) that were originally developed for the WordNet (Miller, 1995) taxonomy, a lexical database for the English language. These adapted measures were tested against sequence similarity using the average combination approach. Later, Pesquita et al. (2008) also tested several measures against sequence similarity and found simGIC to provide overall better results. In contrast, Guo et al. (2006) found simUI to be the weakest measure when evaluated for its ability to characterize human regulatory pathways, while it was found to perform fairly well when evaluated against sequence similarity in the assessment by Pesquita et al. (2008).

2.4. TERM ENRICHMENT

Among the analysis operations involving GO terms, term enrichment analysis is one the most commonly used. Micro-array experiments often output lists which can represent hundred or thousands of genes found to be differentially regulated for a given condition under study. The purpose of term enrichment analysis is then to abstract from the individual genes and focus instead on a representative set of activity terms that summarize the particular biological activity differential, characteristic of the condition being studied. Those differentials (typically enrichment, although it can also be depletion) can be quantitatively measured resorting to commonly used statistical tests for this effect, such as the Fisher exact test, the Chi-squared test, the Hypergeometric distribution and Binomial distribution.

Huang et al. (2009) collected and reviewed 68 bioinformatic enrichment tools categorizing them into three different classes, singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) and modular enrichment analysis (MEA). Common to these three categories is the computation of *p*-values which for SEA is done for each term in a list of pre-selected genes deemed of interest, whereas GSEA needs no pre-selection and has experimental values integrated directly into *p*-value calculation. On the other hand MEA is similar to SEA but additionally factors term-term and gene-gene relations into the *p*-value calculations.

However, and despite the number of available enrichment tools there are still several unaddressed issues, even if we disregard issues stemming from experimental design and execution. These originate from variations in the sizes of the lists of genes, dependencies among genes or terms, annotation incompleteness and overall heterogeneity regarding specificity of annotation. And while the MEA methods try to address and even take advantage of the possible dependencies between genes or terms, issues pertaining to heterogeneous term availability or annotation distribution can still cause several problems and are still not optimally addressed.

3. DISCUSSION

3.1. CASE STUDY

Consider, as case-study, the CAZy database (www.cazy.org) that describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (Cantarel et al., 2009). Its maintenance is done by a small team of curators that uses semi-automatic methods to keep it up-to-date. Even with part of the procedure being automatic there is still a large workload of manual curation that has to be performed by the specialized curators. Recently, the CAZy database has shifted from a schema where function was attributed to the complete enzyme sequence to a schema where function may be assigned just to the segment of the sequence involved in each function, the functional module. So far the CAZy families have been functionally annotated with Enzyme Commission (EC) numbers (Webb and NC-ICBMB, 1992). The EC number is a numerical classification for enzymes, based on the reactions they catalyze. The module-centric organization schema of the database can be complemented in such a way that functions, enzymatic or not, may be directly assigned to a specific segment of a sequence. In summary, CAZy is a curated knowledgebase of functionally related protein (module) families and despite not making use of GO as primary annotation system it still requires annotations with high specificity in order to achieve better characterization. Therefore, the CAZy families are good candidates for performing annotation coherence assessments and annotation extension studies.

The Polysaccharide Lyases (PL) are a group of enzymes that cleave uronic acid-containing polysaccharide chains via a β-elimination mechanism to generate an unsaturated hexenuronic acid residue and a new reducing end. Within the CAZy database these enzymes are classified into families and subfamilies based on amino acid sequence similarities, intended to reflect their structural features (Lombard et al., 2010). A quick assessment of the GO annotation status of these PL families was done using two simple naive metrics, GOscore and GOoccurence (Bastos et al., 2011) described by Equation 2 and Equation 3, respectively.

$$\mathbf{GOscore}(\text{fam}) = \text{MAX}_{\text{term} \in \text{fam}} \left[\text{freq}_{\text{fam}}(\text{term}) \text{x} IC(\text{term})\right] \quad (2)$$

$$\mathbf{GOoccurence}(\text{fam}) = \text{AVG}_{\text{term} \in \text{fam}} \left[\text{freq}_{\text{fam (term)}}\right] \quad (3)$$

Fundamentally, the GOscore metric is an indicator of the maximum *IC* expressed by the annotations of a family as conveyed through the most predominant and most informative term annotating a given family. On the other hand, the GOoccurrence

metric expresses annotation coherence by averaging the frequency of all terms annotating one family. Hence, a family will report maximum functional annotation coherence (GOoccurrence = 1) when all terms are shared by all proteins in a given family. It should be noted that when applying this metric to sets of families of multifunctional proteins misinterpretations can be made if the multiple functions are not evenly shared and annotated within the protein set or family being measured.

3.2. RESULTS

The incompleteness of annotation over a given protein space may lead to erroneous interpretations regarding functional coherence of that space. As mentioned previously we applied two annotation metrics, GOscore and GOoccurence to the PL families of the CAZy database. The results for both metrics are shown in **Table 1** with the respective number of annotated proteins for each family. Upon inspection of the obtained GOoccurrence values, the families PL5, PL15, PL16, PL17, PL20 stand out as being the perfectly coherent families in terms of annotation (GOoccurrence = 1). Further and closer inspection of the actual annotation distribution within those families reveals that families PL5, PL16, PL17 are functionally mono-specific. This means that, for each of those families, there is only a single and common (known) molecular function activity performed by their proteins. Additionally, the reported GOscores for these families are also fairly high and thus indicate that they are annotated with functionally specific terms. Regarding families PL15 and PL20 they present deceptively high GOoccurrence values but these can be dismissed on account of the low number of annotated proteins (3 and 1, respectively) in those families. Given their low statistical support these two families are unsuitable for further analysis. Moreover, the only functional annotation in these two families is the *lyase activity* term. Considering their low *IC* (0.202) the functional information provided by these families is therefore also of little informative value.

In turn, family PL22, despite appearing to be mono-specific, nonetheless has a GOoccurrence score of 0.880. This is in fact due to the penalization inflicted by 7 out of 29 proteins annotated proteins not being annotated with the most specific term *oligogalacturonide lyase activity*. Instead those 7 proteins are only annotated with *lyase activity*, an ancestor term of *oligogalacturonide lyase activity*. So, in this family, despite being mono-specific, it provides a clear case of annotation incompleteness that could lead to misinterpretations if we were to rely on coherence metrics alone. On the other hand, these annotations could be potentially extended to the *oligogalacturonide lyase activity* term. For instance, using all the proteins annotated to this term to create multiple sequence alignments, and subsequently creating position-specific scoring matrices, hidden Markov models or others statistical models these could be used to find matches on the 7 incompletly annotated proteins.

Another example, the PL3 family, despite having a similar GOscore (0.593), conversely has a rather low GOoccurrence score (0.306). In addition, by looking at the distribution of annotation terms within this family we can discover that all of its 228 proteins are annotated to the *pectate lyase activity* term. However, this otherwise coherent annotation is broken by 6 additional terms that annotate the family heterogeneously to a much lesser extent (only up to 2 proteins per term). Thus, here can be seen that the multifunctional nature of proteins can greatly affect the GOoccurrence metric. However, given the context of the PL enzyme class in which the PL3 family is inserted, if we were only to consider annotation terms that are children of *lyase activity* then we would obtain a considerable GOoccurrence improvement to a score of 0.798 (data not shown) for this particular family. The annotation terms that would be discarded, in this case, are clearly the product of secondary functional modules in the proteins that do not contribute to the global functional characterization of the family. Hence, their removal when accounting for family functional coherence is appropriate for this particular case. Regardless of any analitical assertion over their biological value, their low annotation count does not lend additional statistical support. That can be further confirmed through the use of enrichment analysis on this family and using the Benjamini-Yekutieli (Benjamini and Yekuteli., 2001) method, for an $\alpha = 0.01$ only *pectate lyase activity* and *pectin lyase activity* are considered significant (corrected p-values of 0 and 9.8×10^{-4}, respectively).

Visualization can be very helpful when analysing GO term annotations for families or sets of proteins, thus we also used it in our analysis. PL4 is a moderately annotated family in terms of incompleteness which presents low values both for the GOscore and GOoccurrence metrics. The graph represented in **Figure 3** subsumes the GO term annotations from the molecular function aspect in the PL4 family. The top unlabelled node on the graph is actually the root term *molecular_function* to which all the 43 sequences are annotated. It is important to notice that in the graph all indirect or inherited annotations are represented by unlabelled white nodes while direct annotations are represented by gray GO term-named nodes. It should also be noted that the direction of edges on the depicted graph in **Figure 3** is reversed in relation to the actual GO graph. The edges in a typical GO graph represent the hierarchical *is_a* relations that hold between the molecular function aspect terms, and edge direction points from the outer leaf terms converging into a common

Table 1 | GO annotation scores (GOscore and GOoccurrence) and respective size in number of annotated proteins for each CAZy family in the PL enzyme class.

Family	PL1	PL2	PL3	PL4	PL5	PL6	PL7	PL8	PL9	PL10	PL11	PL12	PL13	PL14	PL15	PL16	PL17	PL18	PL20	PL22
Size	391	34	228	43	37	21	63	184	89	77	44	19	5	9	3	22	30	3	1	29
GOocc	0.146	0.798	0.306	0.373	1.000	0.405	0.288	0.303	0.128	0.261	0.325	0.586	0.550	0.420	1.000	1.000	1.000	0.667	1.000	0.880
GOscore	0.196	0.511	0.593	0.309	0.599	0.192	0.202	0.508	0.166	0.202	0.129	0.202	0.718	0.180	0.202	0.640	0.599	0.202	0.202	0.577

root node, making the foundations of the true path rule that states that "the pathway from a child term all the way up to its top-level parent(s) must always be true" (Ashburner et al., 2000). On the other hand, the graph edges on **Figure 3** actually represent the flow of proteins from the most generic root term into the more specific leaf GO terms. Additionally, edge thickness is proportional to the number of proteins "flowing down" from a parent node to a child node, and hence receiving a more specific annotation. Thus, these modified edges are particularly useful in providing visual cues regarding annotation specificity, homogeneity and functional relevance for a given protein family.

Again, given that the PL4 family belongs to the PL enzyme class it would be expected that all proteins within the family might be annotated to the *lyase activity* term. However, out of 43 proteins only 25 are annotated with the *lyase activity* term thereby leaving 18 proteins that potentially could also be annotated with it. By following the descendants of the *lyase activity* term down the graph we find that the term *carbon-oxygen lyase activity, acting on polysaccharides* annotates only 16 sequences. It is not unexpected that the number of annotated protein decreases as we walk down an annotation graph toward the leaf terms. Given that the bulk of annotation is performed by automatic methods it becomes more difficult to provide protein annotations at more functionally specific levels with enough confidence. However, for the PL4 family the most specific GO term is *carboxypeptidase activity* annotating 7 proteins. Although this term is not a descendent of the *lyase activity* term, 6 of the proteins annotated with it are also annotated with with *carbon-oxygen lyase activity, acting on polysaccharides* term. Unlike the PL3 family, for the case of the PL4 family it is not as simple to resolve the multi-functional nature of their proteins and just excluding terms that are not descendants of *lyase activity* is not an obvious option.

InterPro (Zdobnov and Apweiler, 2001) is a resource that can be used to scan protein sequences against an extensive collection of signatures from multiple and diverse databases, and allows the presence of domains and important sites useful to be predicted for protein functional analysis. Therefore, by using the InterProScan on the PL4 family sequences we can obtain the resulting matches against the InterPro signatures. A quick visual comparison of the signature profiles of both *lyase activity* annotated proteins and non-*lyase activity* annotated proteins leads us to infer that the latter can in fact also be annotated to the *lyase activity* term with reasonable confidence given the similarity of the signature profiles. However, as can be seen in **Table 2**, despite the term *lyase activity* being statistically significant, for $\alpha = 0.01$ and a Benjamini-Yekutieli corrected *p*-value, the *IC* (normalized for the GOA annotation corpus) is relatively low, therefore indicative of a differentially low informative value. According to the term enrichment corrected *p*-values, the term *carbohydrate binding* has the greater statistical significance (among all the direct annotations in family PL4). However, intuitively it can be seen that this term, despite being biologically relevant, does not provide a great information increment, since it has the third lowest *IC* value in **Table 2**. The term *carbon-oxygen lyase activity, acting on polysaccharides* ranks second in terms of significance but its *IC* is also only slightly higher than the one for *carbohydrate binding*. It is actually the third ranked term for statistical significance, *carboxypeptidase activity* that has the greatest *IC* even though it is not even a descendant of *lyase activity*. Both *calcium ion binding* and *catalytic activity* fall below the previously chosen threshold of significance. The former can be explained by the fact that it has only one annotation occurrence, and is most likely not relevant for the PL4 family functional profile. The lack of significance of the latter term is explained by its ubiquitousness both within the CAZy families and the GOA annotation corpus which in turn also reflects itself as a low IC.

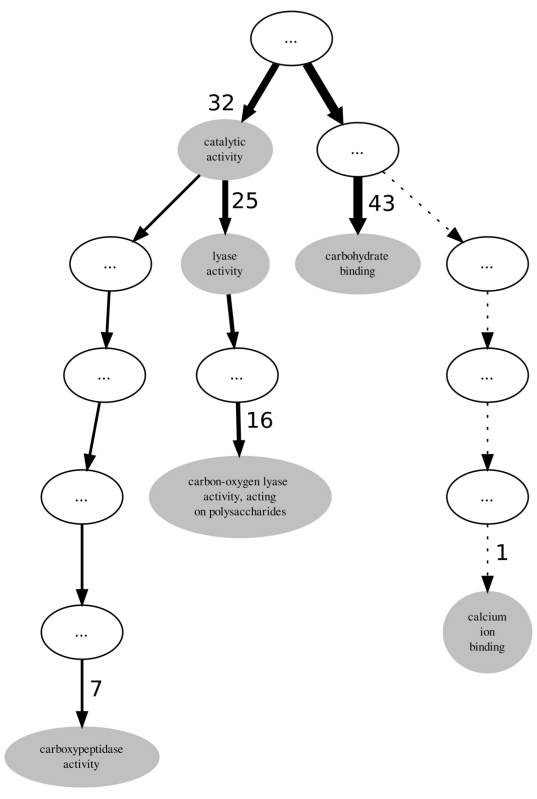


FIGURE 3 | Graph subsuming the GO molecular function aspect annotation of CAZy's PL4 family.

Table 2 | GO term enrichment for CAZy family PL4 with Benjamini-Yekuteli corrected *p*-values, normalized *IC* and number of annotations.

GO term	*p*-value (corr)	IC (norm)	Annotations
Carbohydrate binding	1.87e-47	0.658	43
Carbon-oxygen lyase activity, acting on polysaccharides	7.50e-23	0.699	16
Carboxypeptidase activity	3.43e-12	0.813	7
Lyase activity	1.76e-10	0.404	25
Calcium ion binding	4.75e-01	1.000	1
Catalytic activity	1.00e+00	0.166	32

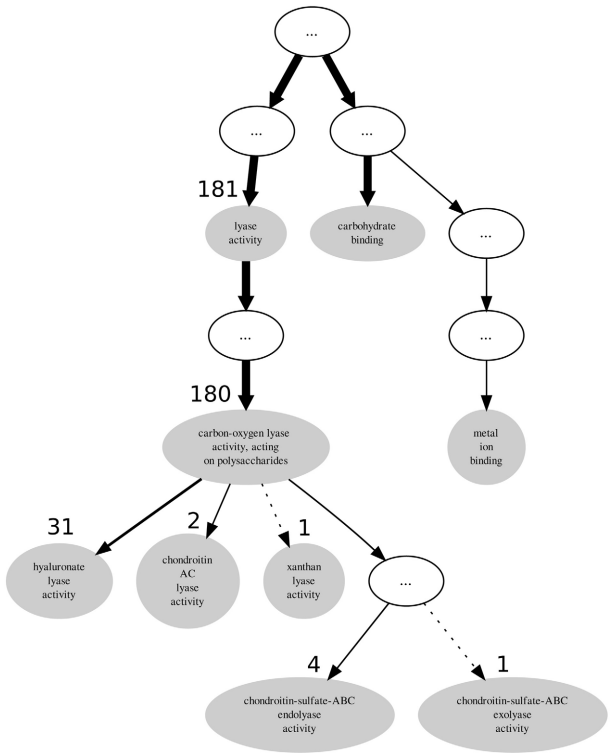


FIGURE 4 | Graph subsuming the GO molecular function aspect annotation of CAZy's PL8 family.

Table 3 | GO term enrichment for CAZy family PL8 with Benjamini-Yekuteli corrected *p*-values, normalized *IC* and number of annotations.

GO term	*p*-value (corr)	IC (norm)	Annotations
Carbon-oxygen lyase activity, acting on polysaccharides	8.15e-306	0.699	180
Carbohydrate binding	1.18e-186	0.658	178
Hyaluronate lyase activity	2.00e-095	1.000	31
Chondroitin-sulfate-ABC endolyase activity	3.69e-011	1.000	4
Chondroitin AC lyase activity	1.24e-005	1.000	2
Chondroitin-sulfate-ABC exolyase activity	5.49e-003	1.000	1
Xanthan lyase activity	5.49e-003	1.000	1
Lyase activity	3.37e-002	0.404	181
Metal ion binding	1.00e+00	0.687	2

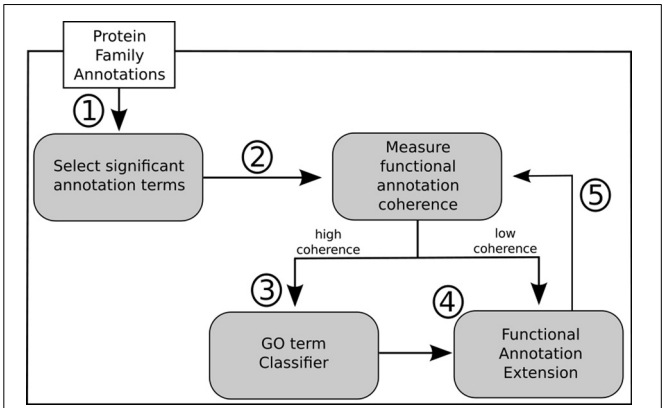


FIGURE 5 | Outline of proposed methodology for annotation extension.

Descendant terms of *lyase activity* can also lead to a reduced annotation coherence, as measured simplistically by the GOoccurrence metric. Any non-uniform annotation distribution within a family will penalize this metric. For the PL8 family, as can be seen in **Figure 4**, the penalization comes in part from the multiple descendants of the *lyase activity* term. There are five leaf-terms that are descendants of *lyase activity* in family PL8, but presenting an asymmetrical distribution regarding the number of proteins they annotate. As show in **Table 3** all of these five terms are enriched in family PL8, however only the term *hyaluronate lyase activity* annotates sufficient proteins to potentially create a support corpus that would allow annotation extension for this term and within this family. Hence, there are still 149 candidate proteins annotated with the *carbon-oxygen lyase activity, acting on polysaccharides* term that can be asserted for extension with the *hyaluronate lyase activity term*. As for the remaining sibling terms they can not be dismissed as irrelevant for the family characterization, and are part of this family set of relevant activities but lowering the value of the GOocccurrence metric.

3.3. PROPOSED APPROACH

In light of the results discussed above we propose a general methodology for extending GO annotations in protein families as depicted in **Figure 5**. Consider a set of protein families created by curators within a given biological knowledge domain. A certain level of functional similarity is inherently expected from these families. Following an initial collection of terms annotating each of these families a statistical enrichment can then ensue. The commonly used technique of statistical enrichment allows the filtering out of possible annotation terms that are not characteristic of a family. At this point (Step 1) additional manually created rules might be beneficial in order to capture not only statistical support but potentially biological meaning related to the specific context domain of the protein families. Following the process of selecting the relevant term annotations for a given family, functional annotation coherence in a family can be asserted through the use of groupwise semantic similarity metrics (Step 2). A protein family showing greater annotation coherence may supply sub-sets of protein (sequences) that can be used to create multiple sequence alignments. These can subsequently be used to create position-specific scoring matrices, hidden Markov models or other statistical models that can be used for classification. Also, any other available or obtainable protein feature from a sub-set of proteins sharing an annotation can theoretically be used with several machine learning techniques in order create individual GO term classifiers. Visualization methods can be helpful in

making this procedure semi-automatic. Following that course of action subsuming annotation graphs, like the ones in **Figures 3**, **4**, can be dynamically generated. These annotation graphs can also be made interactive in order to allow navigation through the individual nodes. Hence, considering that each node represents an annotation term, the graph can then be linked with the subset of proteins annotated by that term in a given family. This allows the selection of proteins which will contribute with features (sequences or otherwise) to construct the single GO term classifiers (Step 3). In turn, these classifiers can then be used for the purpose of extending functional annotation on incompletely annotated proteins within the given protein family (Step 4). By submitting the families to the annotation metrics the coherence differential can be gauged after each iteration of annotation extension (Step 5). It should be noted that the overall family coherence metrics used should be selected or customized in order to take into account the particular knowledge domain being assessed. Of particular notice is that extensions are done per annotation term, and each protein (and family) can have multiple functions and thus terms associated to them.

4. CONCLUSIONS

Ideally, proteins should be annotated in a way that fully describes their functional activities. However, even within the boundaries of current knowledge, this is seldom the case. As we try to compare protein sets, such as families, based on their functional annotations this heterogeneity of functional annotation becomes a greater issue. Annotation incompleteness in annotations can lead to false interpretations about the existing functional inter-similarity within a given protein set (or family). In order to avoid erroneous interpretations on heterogeneous protein sets or families (in terms of annotation specificity), functional comparisons are usually done at conservative levels. This means that by comparing families at conservative annotation levels we would also be comparing terms with lower *IC* and hence obtaining less informative conclusions.

Resources such as the CAZy database provide high-quality classifications of segments of the protein space into functionally related families. These kind of protein families present themselves as an opportunity and a knowledgebase from which we can benefit in order to provide annotation extension methodologies. Considering that any given protein family is a functionally related set of sequences, then the heterogeneity of annotation specificity can be explored within each family. Thus, sub-sets of homogeneous annotation in a family can be used to produce classifiers which can potentially extend other proteins within the same family that are under-annotated. This proposed methodology should be regarded as a generic approach guided at mitigating some of the current issues with annotation incompleteness, and despite not being suitable for all annotation incompleteness states it should allow for an increased extension of annotation over the ever increasing protein space. It should be particularly useful when applied in tandem with protein families from databases like CAZy where proteins despite being grouped together into functionally close families they still do not focus on functional annotation. It should be noted that the coherence metrics presented here are only to illustrate typical annotation baseline patterns and are not intended to be used in fully automated procedures or to address issues like the measuring of coherence in sets of multifunctional proteins on their own. However, customized metrics derived from groupwise semantic similarity measures can be implemented for each specific knowledge domain under study in order to automate most of the procedure in the suggested methodology.

ACKNOWLEDGMENTS AND FUNDING

Portuguese Fundacão para a Ciência e Tecnologia (www.fct.pt/) through the financial support of the SPNet project (PTDC/EIA-EIA/119119/2010), the SOMER project (PTDC/EIA-EIA/119119/2010) and the PhD Grant ref. SFRH/BD/48035/2008 and through the LASIGE multi-annual support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2013.00201/abstract

REFERENCES

Alterovitz, G., Xiang, M., Hill, D. P., Lomax, J., Liu, J., Cherkassky, M., et al. (2010). Ontology engineering. *Nat. Biotechnol.* 28, 128–130. doi: 10.1038/nbt0210-128

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic. Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Attwood, T. K. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31, 400–402. doi: 10.1093/nar/gkg030

Barrell, D., Dimmer, E., Huntley, R. P., Binns D, O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* 37, D396–D403.

Bastos, H., Couto, F., and Coutinho, P. (2011). "Exploring gene ontology relationships in enzyme families: an application to polysaccharide lyases," in *9th Carbohydrate Bioengineering Meeting (CBM9)*, (Lisbon).

Benjamini, Y., and Yekutieli, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* 33, D212–D215. doi: 10.1093/nar/gki034

Bowers, P., Pellegrini, M., Thompson, M., Fierro, J., Yeates, T., and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* 5:R35. doi: 10.1186/gb-2004-5-5-r35

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37, D233–D238. doi: 10.1093/nar/gkn663

Chua, H. N., Sung, W. K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22, 1623–1630. doi: 10.1093/bioinformatics/btl145

Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. (2002). Prediction of protein function using protein-protein interaction data. *Proc. IEEE Comput. Soc. Bioinform. Conf.* 1, 197–206. doi: 10.1109/CSB.2002.1039342

Enault, F., Suhre, K., and Claverie, J. M. (2005). Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis.

BMC Bioinformatics 6:247. doi: 10.1186/1471-2105-6-247

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., et al. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222. doi: 10.1093/nar/gkp985

Gan, M., Dou, X., and Jiang, R. (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity *Scientific World Journal* 2013:793091. doi: 10.1155/2013/793091

Gentleman, R. (2005). *Visualizing and Distances Using GO*. Available online at: http://www.bioconductor.org/docs/vignettes.html

Gonzalez, O., and Zimmer, R. (2008). Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. *Bioinformatics* 24, 1257–1263. doi: 10.1093/bioinformatics/btn106

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 22, 967–973. doi: 10.1093/bioinformatics/btl042

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Jaccard, P. (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37, 547–579.

Jiang, J., and Conrath, D. (1997). "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the International Conference on Research in Computational Linguistics*, (Taiwan), 19–33.

Letovsky, S., and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 19, 197i–204i. doi: 10.1093/bioinformatics/btg1026

Lin, D. (1998). "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, (Madison), 296–304.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., and Henrissat, B. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J.* 432, 437–444. doi: 10.1042/BJ20101185

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275–1283. doi: 10.1093/bioinformatics/btg153

Martin, D. M. A., Berriman, M., and Barton, G. J. (2004). GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* 5:178. doi: 10.1186/1471-2105-5-178

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad Sci. U.S.A.* 96, 2896–2901.

Pesquita, C., Faria, D., Bastos, H. P., Ferreira, A. E. N., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* 9(Suppl 5):S4. doi: 10.1186/1471-2105-9-S5-S4

Pesquita, C., Faria, D., Falcão A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* 5:e1000443. doi: 10.1371/journal.pcbi.1000443

Pesquita, C., and Couto, F. M. (2012). Predicting the extension of biomedical ontologies *PLOS Comput. Biol.* 9:e1002630. doi: 10.1371/journal.pcbi.1002630

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man. Cybernet.* 19, 17–30. doi: 10.1109/21.24528

Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th international joint conference on Artificial intelligence - IJCAI'95)*, Vol. 1, (Montreal), 448–453.

Rost, B. (2002). Enzyme function less conserved than anticipated. *J. Mol. Biol.* 318, 595–608. doi: 10.1016/S0022-2836(02)00016-5

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 38, D161–D166. doi: 10.1093/nar/gkp885

Sleator, R. D., and Walsh, P. (2010). An overview of *in silico* protein function predition. *Arch. Microbiol.* 192, 151–155. doi: 10.1007/s00203-010-0549-9

Vinayagam, A., del Val, C., Schubert, F., Eils, R., Glatting, K. H., Suhai, S., et al. (2006). GOPET: a tool for automated predictions of Gene Ontology terms. *BMC Bioinformatics* 7:161. doi: 10.1186/1471-2105-7-161

Webb, E., and NC-ICBMB. (1992). *Enzyme Nomenclature 1992: Recommendations of the NCIUBMB on the Nomenclature and Classification of Enzymes, Enzyme Nomenclature.* San Diego, CA: Academic Press. doi: 10.1109/TSP.2006.873718

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zhang, X., Kim, S., Wang, T., and Baral, C. (2006). Joint learning of logic relationships for studying protein function using phylogenetic profiles and the rosetta stone method. *IEEE Trans. Signal Process.* 54, 2427–2435.

Zhang, W., Morris, Q., Chang, R., Shai, O., Bakowski, M., Mitsakakis, N., et al. (2004). The functional landscape of mouse gene expression. *J. Biol.* 3, 21. doi: 10.1186/jbiol16

Zheng, Y., Roberts, R. J., and Kasif, S. (2002). Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.* 3, 0060.1–0060.9. doi: 10.1186/gb-2002-3-11-research0060

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 June 2013; accepted: 22 September 2013; published online: 11 October 2013.

*Citation: Bastos HP, Clarke LA and Couto FM (2013) Annotation extension through protein family annotation coherence metrics. Front. Genet. **4**:201. doi: 10.3389/fgene.2013.00201*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

METHODS ARTICLE
published: 05 October 2012
doi: 10.3389/fgene.2012.00197

The ontology-based answers (OBA) service: a connector for embedded usage of ontologies in applications

***Jürgen Dönitz*[1,2]* and *Edgar Wingender*[1]**

[1] Department of Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[2] Department of Developmental Biology, Johann-Friedrich-Blumenbach Institute for Zoology and Anthropology, Georg-August University Göttingen, Göttingen, Germany

Edited by:
John Hancock, Medical Research Council, UK

Reviewed by:
Katy Wolstencroft, University of Manchester, UK
Damian Smedley, Sanger Institute, UK

***Correspondence:**
Jürgen Dönitz, Department of Bioinformatics, University Medical Center Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, Germany.
e-mail: juergen.doenitz@bioinf.med.uni-goettingen.de

The semantic web depends on the use of ontologies to let electronic systems interpret contextual information. Optimally, the handling and access of ontologies should be completely transparent to the user. As a means to this end, we have developed a service that attempts to bridge the gap between experts in a certain knowledge domain, ontologists, and application developers. The ontology-based answers (OBA) service introduced here can be embedded into custom applications to grant access to the classes of ontologies and their relations as most important structural features as well as to information encoded in the relations between ontology classes. Thus computational biologists can benefit from ontologies without detailed knowledge about the respective ontology. The content of ontologies is mapped to a graph of connected objects which is compatible to the object-oriented programming style in Java. Semantic functions implement knowledge about the complex semantics of an ontology beyond the class hierarchy and "partOf" relations. By using these OBA functions an application can, for example, provide a semantic search function, or (in the examples outlined) map an anatomical structure to the organs it belongs to. The semantic functions relieve the application developer from the necessity of acquiring in-depth knowledge about the semantics and curation guidelines of the used ontologies by implementing the required knowledge. The architecture of the OBA service encapsulates the logic to process ontologies in order to achieve a separation from the application logic. A public server with the current plugins is available and can be used with the provided connector in a custom application in scenarios analogous to the presented use cases. The server and the client are freely available if a project requires the use of custom plugins or non-public ontologies. The OBA service and further documentation is available at http://www.bioinf.med.uni-goettingen.de/projects/oba

Keywords: ontology, semantic function, ontology-based answers, OBA

INTRODUCTION

Ontologies play a major role in the semantic web (Berners-Lee et al., 2001). Running in the background they provide electronic systems with the expertise of a knowledge domain. Through formal and logical statements ontologies are useful to unambiguously identify and define entities representing material objects as well as abstract concepts and their mutual relations. By connecting unknown terms with known ones through defined statements, new knowledge can be deduced. This knowledge can be used to provide the user with information that he/she is seeking but could not exactly specify. This is achieved by means of a mandatory class hierarchy, using the "is_a" relation, and other relations, connecting the ontology classes to each other. Supplementary data can be added to each ontology class by annotations. While the meaning of relations is comprehensible to human users so that they can select the right one for traversing the graph, it is a particular challenge to transfer the logical axioms defined in an ontology into an object-oriented view that is common to most applications (Winston et al., 1987; Burger et al., 2008).

A multitude of tools and web services dealing with ontologies are available in the biomedical field. Ontology browsers like Amigo (Carbon et al., 2009) for the Gene Ontology (GO; Ashburner et al., 2000) or ontology editors (OBOEdit: Day-Richter et al., 2007; Protégé[1]) let the user work interactively with an ontology. The web services Ontology Lookup Service (OLS; Côté et al., 2008), the NCBO BioPortal (Noy et al., 2009) and OntoCAT (Adamusiak et al., 2011) facilitate the search function covering all ontologies publicly available at the NCBO portal or the OBO-Foundry (Smith et al., 2007) and provide access to their content. OntoCAT and the BioPortal also offer an interface to be queried by electronic systems over the network. By doing so OntoCAT additionally offers a Java and R client (Kurbatova et al., 2011) for communication with the service.

The listed portals offer services which are highly valuable to the community. However, they fall short in two aspects: by approaching the access of a collection of ontologies in a standardized way, the portals lack functions that are specific for individual ontologies, leaving the information encoded in the diverse relationships unattended. An automated system does not allow the user to decide when to use which relationship, the algorithm

[1] http://protege.stanford.edu

has to solve this problem. The application developer is required to be familiar with the annotation guidelines and implement the required algorithm.

If a search interface allows the user to enter or select an anatomical structure, for which data should be displayed, the user will expect results not only for the selected structures, but also for substructures and perhaps functionally related structures. With the use of ontologies this challenge can be met. The different sets of available relations used in ontologies like "part_of," "contained," or "bordered_by" require an implementation of such a search algorithm to be ontology specific.

The other challenge is between the semantics of ontologies, consisting of a set of axioms, and the modern style of object-oriented programming. In an ontology the classes and their relations are stored in separate axioms while in an object graph the objects themselves have knowledge about the links to their neighbors. APIs like OWL-API (Horridge and Bechhofer, 2011) or Jena-API (Jena – A Semantic Web Framework for Java[2]) facilitate full access to ontologies and follow their design principles. They disclose any information and logic of the supported ontology format to the user. The resulting complexity prevents a straight way to get, e.g., neighbors of a class from the ontology. To get the subclasses of an ontology class with the OWL-API the axioms for the superclass has to be fetched and the right axioms have to be selected. Also when using the ontology portals an additional request to the portal is required because the ontology classes fetched from the portals lack a method to access their own subclasses.

As an alternative way we suggest a service providing ontology-based answers (OBA service). To benefit from ontologies the OBA service can be embedded in applications and workflows. The OBA project's goal is to make knowledge, which a user can intuitively retrieve from ontologies, available to applications or to workflows processing high-throughput data. The service provides semantic functions that implement knowledge about the curation guidelines as well as the used relations and their interpretation. The client of the service can be embedded into custom applications and maps the service's responses to a graph of Java objects. The OBA service provides the main information stored in ontologies to computational biologist not familiar with ontologies. The developers are enabled to concentrate on their research topic while working with the familiar object-oriented programming style.

Use cases and projects are presented to demonstrate the concept and advantages of OBA. In the use cases the Cytomer ontology and the iBeetle project are used. Cytomer[3] is an ontology concerning anatomical structures of humans in adults and during the fetal development (Heinemeyer et al., 1999; Michael et al., 2005). Specific relations describe the progenitor, the derivation and the appearance in the Carnegie stages.

The iBeetle project[4] aims to identify genes essential to insect development and physiology by genome wide gene silencing in the red flour beetle *Tribolium castaneum* (Schröder et al., 2008) using parental and larval RNA interference (Bucher et al., 2002; Tomoyasu and Denell, 2004). During the first part of the iBeetle project, several thousand genes have been silenced and the observed phenotypes are stored in a database and linked to an anatomical ontology for *Tribolium* (Bucher and Klinger, personal communication).

MATERIALS AND METHODS

A service which helps to bridge the shortcomings of existing tools, as it is described in Section "Introduction," should fulfill the following requirements:

- The service should enable an application developer to deal with the ontology in a transparent manner rather than enforcing him to deal with different ontology formats or low level APIs.
- The service should map the ontology classes and their connections to a graph consisting of Java objects.
- The part processing the ontologies should be separated from the part which is embedded in the application. A server process would in addition offer a central ontology server.
- The communication with the server should be encapsulated by a connector on the client side to provide network transparency for the custom application.
- The service should implement knowledge about the used ontologies and provide the information deduced from the ontologies by simple Java methods to a computational biologist.
- With more in-depth knowledge about the used network interface or ontologies the service should be extensible to match the requirements of new or custom ontologies and projects.

The OBA service consists of a server and a client part, which communicate using the Representational State Transfer (REST) architecture (Fielding, 2000). **Figure 1** gives an overview of the OBA service design. The server can load any ontology in the OWL (Lacy, 2005) or OBO format (Smith et al., 2007) and host semantic functions. For every ontology a basic part of the server provides access to the entities, connected entities and lists of entities. Each entity is accessed by a unique Uniform Resource Locator (URL). Entities linked to another entity, like its child or parent classes, can also be accessed by a URL denoting the required subresource. Like the content of the ontologies, the semantic functions are available through URLs and return entities or a list of entities as answer.

A list of entities can be stored on the server in order to facilitate the work on more comprehensive input. This data can be used, for example, to limit the results of a search to members of a list of entities used in an application. To manage resource allocation, the storage area is divided into partitions. A user or a work group can create their own partition to store one or more lists. Such a partition is only accessible through its assigned name, allowing a basic access control.

The server uses a REST interface and provides the data in the "application/json," "text/plain," and "text/html" format (MIME-types). The open architecture allows the user to communicate with the server via a command line client, a web browser or with any custom client. The preferred form is the embedding in custom applications. For easy integration into applications a Java client is provided. The client encapsulates the network communication and facilitates access to the semantic functions of the server and to the entities of the respective ontology by Java functions.

[2]http://jena.sourceforge.net

[3]http://cytomer.bioinf.med.uni-goettingen.de

[4]http://ibeetle-base.uni-goettingen.de/

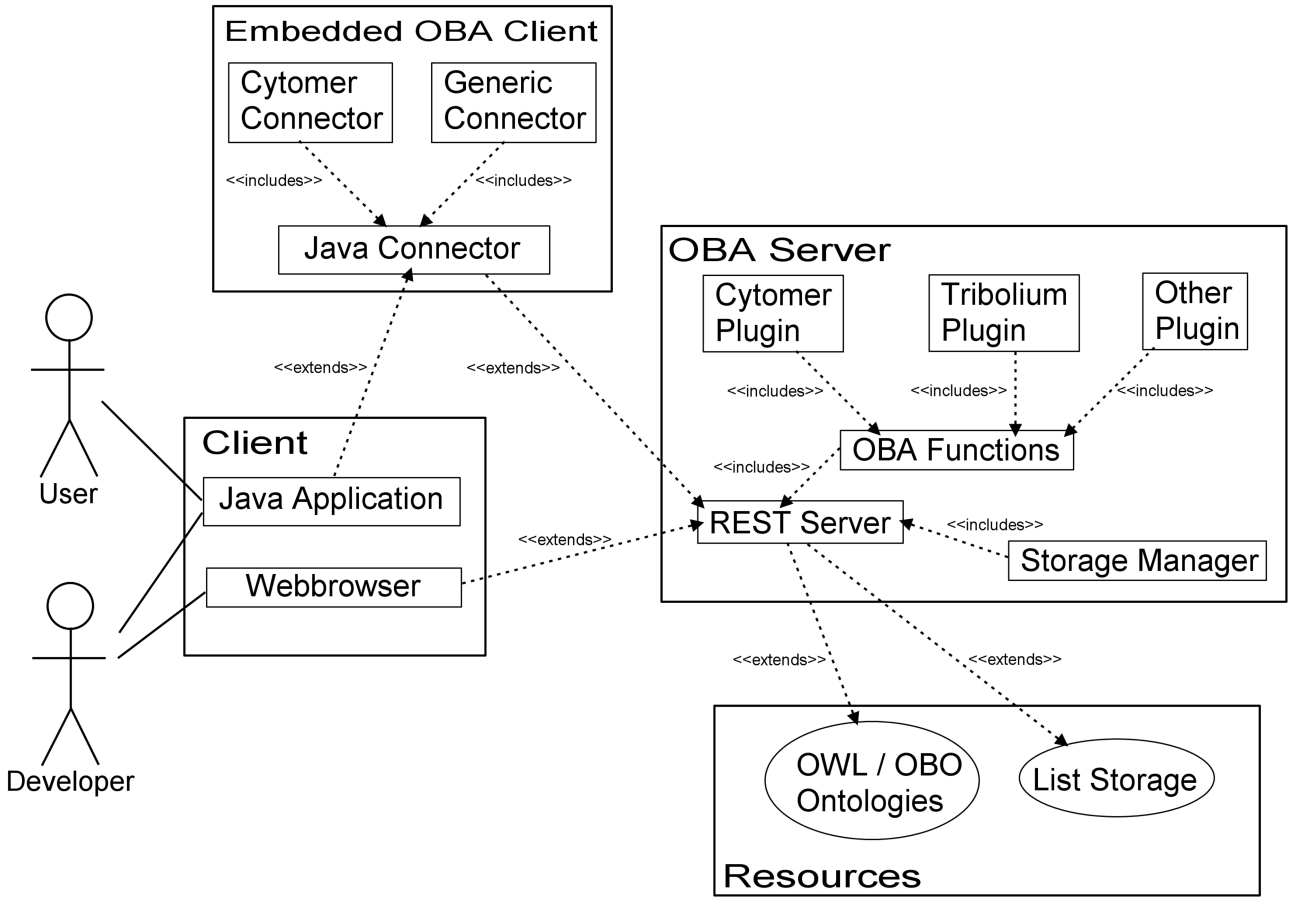


FIGURE 1 | The components of the OBA project and their relations. The components of the OBA project are displayed together with their connections to external parts, to ontology files or to other applications.

The server's response is converted into Java objects, containing methods to access super- and subclasses as well as annotations and relations. To avoid loading the whole ontology upon the first request, the Java objects representing the ontology classes function as proxies that load connected objects upon the first access. This lazy loading is completely transparent to the application.

By default, the client uses the public server available at http://oba.sybig.de. Currently, this server provides access to the Cytomer ontology, the *Tribolium* anatomical ontology (TrOn) and the GO with ontology specific functions for the first two and generic semantic functions for all ontologies. To access custom ontologies or to implement individual OBA functions, the server and the client can be downloaded and extended. The server can load plugins to add custom OBA functions to meet new requirements of a specific project or ontology. The module containing the basic functions implements the plugin interface and can be deemed as built-in plugin. Two additional plugins, one for the Cytomer ontology and one for the iBeetle project, are already available and can serve as templates for the development of new plugins. Client extension is achieved by subclassing. These subclasses can provide Java functions to access semantic functions of a custom plugin or provide convenient functions to access annotations or relations of the ontology's classes. Each ontology has its own defined set of relations and annotations. The generic client has no knowledge of the specific sets of annotations and relations for an ontology and enables access to the annotation and relations as two-dimensional lists containing the type of the annotation or relation and the respective values. To get the synonyms annotations of an ontology class, the application has to iterate the list of annotations until the desired one is found. A custom client can provide the method "getSynonyms()" encapsulating this iteration.

The OBA server and the example client are implemented using the Java Platform. The OWL-API is used to access ontologies in OBO or OWL format. To implement the REST-protocol the Jersey library was selected[5]. The Grizzly HTTP container handles the network communication on the server side[6]. To index the ontology's classes the Lucene library[7] is used. To store the metadata of the uploaded data HSQLDB[8] was selected.

[5]http://jersey.java.net/
[6]http://grizzly.java.net/
[7]http://lucene.apache.org/
[8]http://hsqldb.org/

RESULTS

With the OBA service a software application was developed to fulfill the requirements listed above (see Materials and Methods). The division into a server and a client component allows the separation of processing the ontologies from the specific custom applications. The server has access to the ontologies and hosts plugins with the OBA functions. These functions make intensive use of the ontologies and transfer the processed results to the client. The plugins encapsulate the implementation details to process the ontologies and reduce the complexity to a single function call on the client's side. The concept of the OBA functions as a server side component is a new concept not known to the existing ontology portals.

The OBA client maps the OBA functions to Java functions and the ontology classes to Java objects. The objects representing the ontology classes have functions implemented to access their parents, children, and connected ontology classes. To avoid loading the complete ontology from the very beginning the neighboring classes are loaded upon the first access by a proxy functionality. The Java objects created by the OBA client are internally equipped with a link to the Java connector to load missing neighboring classes in the background. In contrast to existing solutions this loading process is completely transparent to the user. The developer is able to accesses the neighboring classes through Java methods and does not have to be concerned about their loading from the backend. The OBA client facilitates also access to the OBA functions by simple Java methods. Using the OBA client the network access and the implementation details of the OBA functions are transparent to the application developer, who can thus focus on the scope of his custom application.

The following use cases illustrate some OBA functions and how OBA is already used in some upcoming projects. The description of the OBA functions reveals the implementation details of these functions to show how the ontology is processed. The application developer can use these functions with a single function call and is not required to reimplement the logic.

OBA FUNCTION: GENERIC SEARCH

The function "searchCls" is used to search for an ontology class matching a pattern that has been specified by the user. The search is not limited to the name of the ontology class, but the annotation fields of the class are included. On the client side the annotation fields to be used for the search can be specified.

Table 1 shows the result of a search for "cistern" in the Cytomer ontology. In the second case the search is restricted to the annotation "definitionEnglish." The search function of the Java client also provides the possibility of limiting the search to selected annotation fields. This possibility is not common in existing tools but is a powerful filter to get more precise search results.

The search functionality uses the name of the ontology class as well as its annotation fields and works with any loaded ontology. The classes returned by the search function can serve as starting point for traversing the graph or as input for other OBA functions.

OBA FUNCTION: MAP ONTOLOGY CLASSES TO ANCESTORS

The goal of the following two functions is to map ontology classes to more abstract ancestors. The function "reduceToLevel" requires the input of a level and a single ontology class or a list of them. Each one of the classes from the input is mapped to all ancestors at the given level beneath the root node. To determine the ancestors of a class, all paths between the start class and the root class are considered. Due to the fact that an ontology class can have more than one parent, there might be more than one path, resulting in multiple ancestors for a single class at a specific level. If the node "negative regulation of binding" in **Figure 2** is mapped to level five, the two nodes "negative regulation of molecular function" and "regulation of binding" are returned. The function can also be called with a reference to a previously uploaded list of ontology classes. In doing so a list of classes with different levels of abstraction are mapped to classes at a constant and equal level below the root node.

A similar approach is implemented in the function "reduceToClusterSize." In this case the ontology classes are successively mapped to their parents. In each iteration only those classes with the greatest distance to the root class are mapped to their parents. The process is finished when the number of resulting ontology classes is not larger than the specified number. The result is a list of clusters, each with a list of ontology classes from the input list, mapped to this class. Due to the specification of a maximum number of clusters instead of a concrete level, the resulting clusters may have varying distances to the root class. However, by processing the farthest ontology classes in each step, this effect is minimized. The marked nodes in the example of **Figure 2** will be mapped to the nodes "regulation of signaling" and "regulation of protein binding" if the maximum number of clusters is set to the value of two. The node "regulation of cytokine activity" is mapped in

Table 1 | Generic search with a limitation to an annotation field.

http://oba.sybig.de/cytomer/functions/basic/searchCls/cistern	http://oba.sybig.de/cytomer/functions/basic/searchCls;field=definitionEnglish/cistern
cistern, pontocerebellar_cistern, chyle_cistern, ambient_cistern, lumbar_cistern, quadrigeminal_cistern, interpeduncular_cistern, chiasmatic_cistern, pericallosal_cistern, cistern_of_lamina_terminalis, lateral_cerebellomedullary_cistern, vein_of_cerebellomedullary_cistern, posterior_cerebellomedullary_cistern, cistern_of_lateral_cerebral_fossa, basilar_artery	pontocerebellar_cistern, basilar_artery

Result of a search for "cistern" in the Cytomer ontology. In the first case the pattern is searched in the class name and all annotation fields including the comment field. In the right column the search is limited to the annotation "definitionEnglish" by a matrix parameter in the URL.

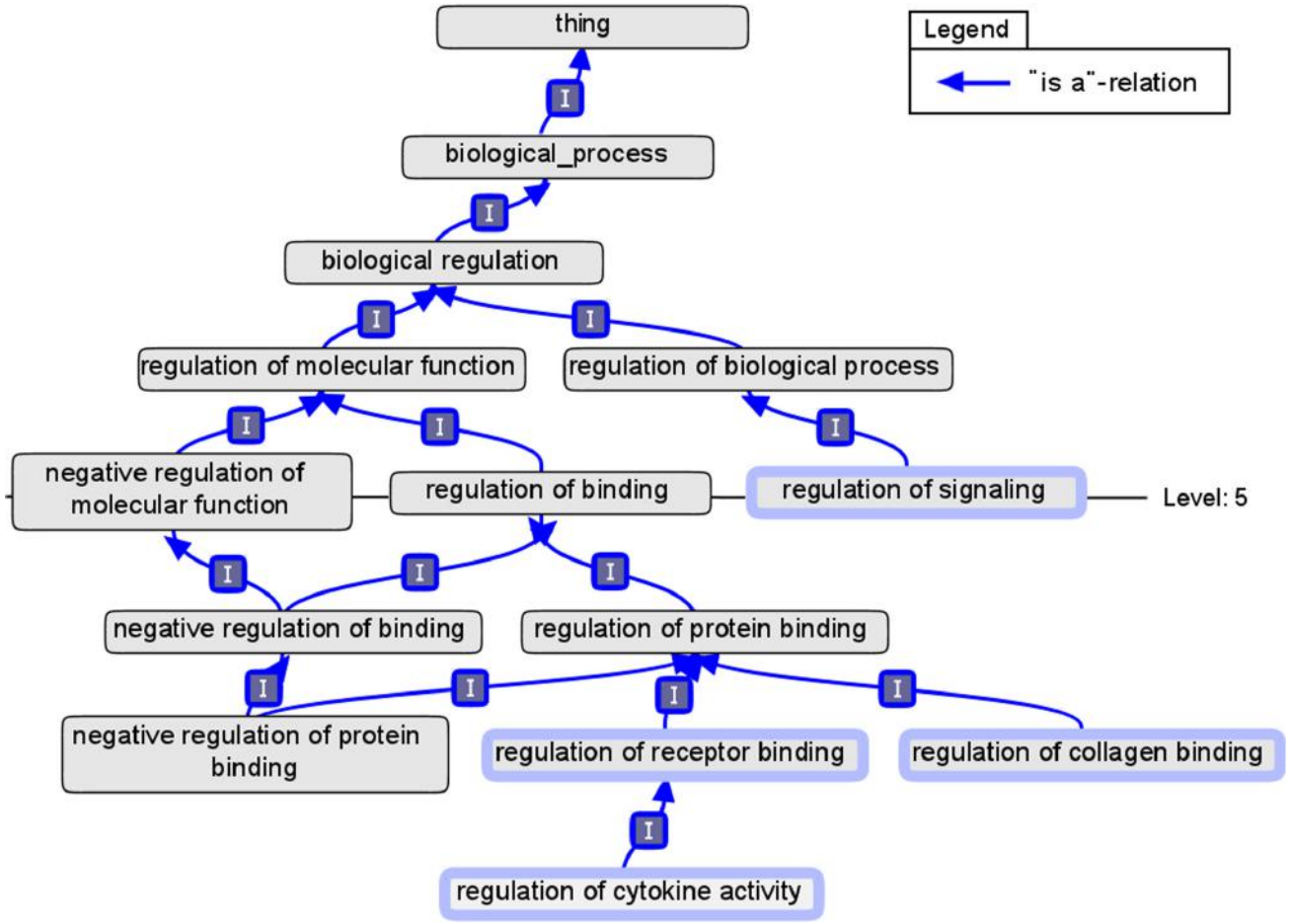


FIGURE 2 | A modified screenshot of the Gene Ontology using OBOEdit. The node "thing" was added as root of the ontology and marks the first level. The marked nodes have a thicker border.

each step, while "regulation of signaling" is just copied to the result set. The classes representing the final cluster do also have different distances to the root node, five and six in this case.

The functions described in this section relay on the class hierarchy and are therefore not ontology specific, they can process any currently loaded ontology as well as the ontologies added in the future. When the described function has to be implemented with existing tools the effort is larger. To map ontology classes to a given level all classes from the starting class up to the root node have to be fetched to determine the classes on the required level. The other classes can be dismissed afterward. The OBA functions simplify the tasks by hiding the processing step behind a function call provided by the OBA client.

The result of a gene expression experiment is a list of differentially expressed genes. A common way to analyze this gene list is to map the genes to the corresponding terms of the GO. The mapping can be done for example with the help of BioMart from Ensembl (Kinsella et al., 2011). Apart from a statistical analysis the resulting list of GO terms can be mapped to more abstract terms until the list is short enough to give an overview of the main processes the GO terms belong to. This can easily be achieved with the two OBA functions "reduceToClusterSize" and "reduceToLevel" and gives a first and intuitive impression of the experiment's outcome.

USE CASE: CYTOMER-SPECIFIC FUNCTIONS

In the following the advantages of OBA functions provided by the service are demonstrated using the anatomical ontology Cytomer. In biomedical research different anatomical structures are investigated. These anatomical structures can be cells, tissues, organs, and entire body parts. A common example is the handling of gene or protein expression data derived from cells, organs, or biopsies (Uhlen et al., 2010). For an analysis on an equal level of abstraction, it is preferable to map all anatomical structures to the level of organs. These steps need to be automated for high-throughput data.

OBA function: get organs of an anatomical entity

The function "organsOf" of the OBA service accepts an arbitrary class of the Cytomer ontology, which represents an anatomical structure as input and returns its respective organs. Inside this function the organs are searched along the class hierarchy and along the selected relations "isPartOf," "isPartOfOrgan," and

"isCellOf." Other relations, for example relations describing the development, are ignored in this case. **Figure 3** shows a simplified, abstract section of Cytomer. Using the function "organsOf" on "Cell 1" "Organ 3" is found using the two relations "isPartOf" and "isCellOf." For "Cell 2" the two nodes "Organ 1" and "Organ 2" are found. "Organ 3" is not part of the result, because the relation "differentiatesInto" between the nodes "Cell 2" and "Cell 1" is not considered for the search of the organs of an anatomical entity. To retrieve the physiological system of an anatomical entity the function "physiologicalSystemsOf" can be used, which works in an analogous way.

OBA FUNCTION: MAP TO A PREDEFINED LIST

An alternative approach is to store the data linked to the most precise anatomical entities, even if these entities do not belong to the same level. In this case a user needs help to draft a request. If the request is on another level than the stored data, no match may be found, although there are relevant entries on a more abstract or more concrete level. The two functions "findUpstreamInSet" and "findDownstreamInSet" of the OBA service provide a solution for this use case. In a set-up step the list of anatomical structures represented by the input data, is stored on the OBA server. The list can be reused for each user's request. Starting from the class, which has been requested by the user, the ontology is searched until a class in the previously uploaded set is found. For an illustration of these two functions please refer to **Figure 4**. The graph is a simplified view of the Cytomer ontology. The yellow nodes are anatomical structures used in EndoNet and uploaded to the OBA service. In the first example the user is searching information on nephron, which would give no result in EndoNet. The function "findUpStreamInSet" searches upstream of the start class "nephron," until a class is found which is also in the previously uploaded list. In this case, following the "isPartOf" relation "kidney" is found, to which EndoNet can provide information to the user. The example of the function "findDownStreamInSet" starts with the abstract term "digestive_organ" and returns "liver" and "pancreas" as matching classes in EndoNet, by following the class hierarchy. The nodes and edges marked with a green shape are the entities processed during the mapping. The search only stops when a member of the predefined list is found, or no more nodes up- or downstream along the class hierarchy or the used relations are available.

These two functions contain a list of relations usable for the up- and downstream search. The path from the starting class to the result nodes may contain any mixture of the intended relations for the requested search direction. The length of the path is not limited, the breadth-first search stops in the iteration step with the first match and returns all matches found in this step.

The OBA functions presented above process a graph's representation of the Cytomer ontology containing the ontology classes, the class hierarchy, and other relationships between the classes. As ontology specific information the functions have the knowledge implemented when to use which relation and how organs or physiological systems can be identified. Processing the graph's

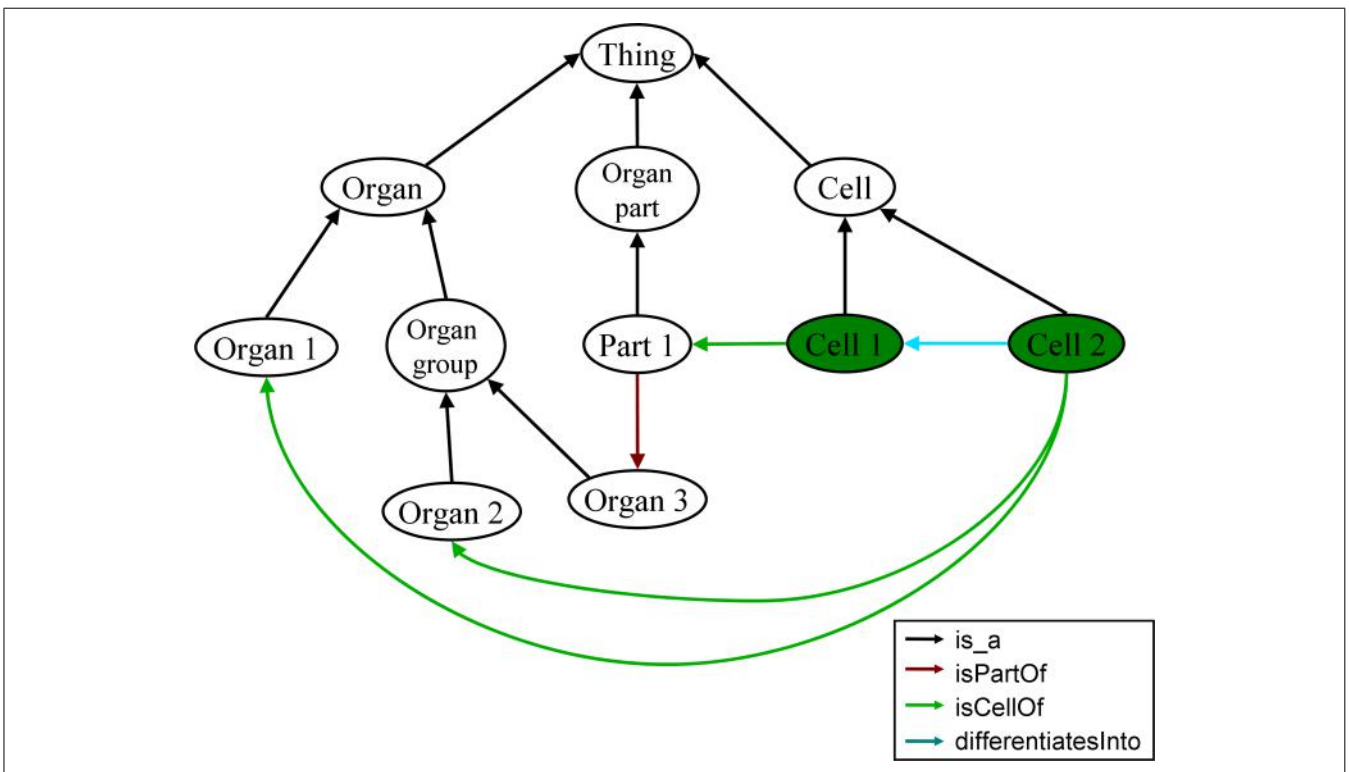


FIGURE 3 | Abstracted and simplified view of the Cytomer ontology illustrating the handling of organs of an anatomical entity. The green nodes are the start nodes for the search function specified in the text. In this section, the entities are connected by four different relations given in the legend.

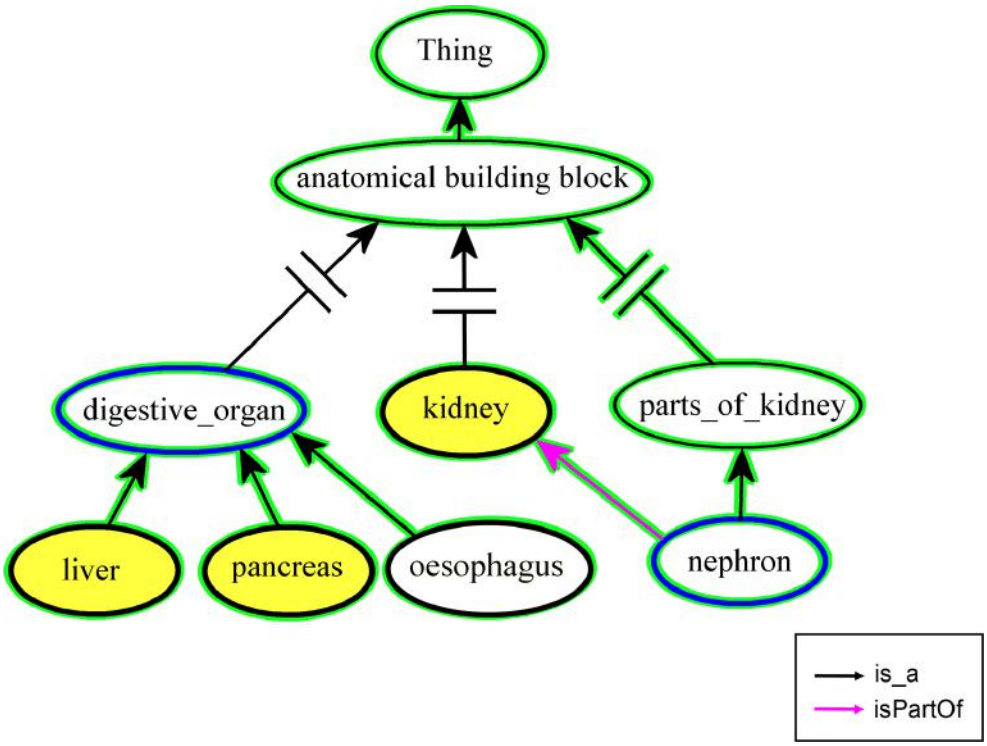


FIGURE 4 | Mapping entities to a predefined list. The nodes with the blue border represent the start nodes for the functions "findUpstreamInSet" and "findDownstreamInSet," respectively. The nodes and edges marked with a green background shape are processed during the mapping. The classes of the result set have to be members of the predefined set which contains the yellow nodes. A predefined list can be used by a project to limit the result of a up- or downstream search to a set of classes used in the project.

representation is done by the OBA framework, to implement analogous functions for other ontologies or similar tasks, a new plugin can reuse this existing logic and only the ontology or task specific knowledge needs to be added, i.e., the relations to use and the key classes.

To achieve a comparable result with existing ontology portals is much more complex. In order to retrieve all organs for an arbitrary anatomical structure using the existing ontology portals the user has to decide which of the relations of the starting class could be used to traverse the ontology graph to some organ. In the next step, all neighboring classes linked by the selected relations have to be queried from the portal. The last two steps have to be repeated for every fetched intermediate class multiplying the number of classes in each step. Whether one of the processed classes represents an organ has to be decided by the users based on their medical knowledge or based on rules deduced from the curation guideline of the ontology. Using the OBA function "organsOf" all these steps are executed on the server where the knowledge is implemented which ontology classes represent the concrete organs. Due to the multitude of relations to consider, 70 ontology classes are processed to return "liver" as organ for the ontology class "hepatocyte." To get the organs lung, larynx, and trachea for the ontology class "sensory_epithelial_cell" 2,497classes are needed to be checked. Without OBA each of these classes has to be downloaded from an ontology portal and processed locally. The numbers are dependent on the starting class and the version of the used ontology. New or removed relations can have a great impact on the number of processed ontology classes. However, for simple queries like the example of the hepatocyte cell, a considerable number of ontology classes already have to be processed. Using OBA the result is always achievable with one single function call. Even changes in the annotation guidelines, like new relations' types, of the used ontology would be encapsulated in the plugin and hidden from the application developer.

PROJECT: iBeetle

In the iBeetle project genes are silenced by RNAi and the observed phenotypes for several stages are annotated into a database following the Entity–Quality (EQ) system (Washington et al., 2009). During the project a detailed ontology about the anatomical structures of *Tribolium* in different developmental stages has been created. There is an ontology class for each structure at every developmental stage where this structure exists. Thus there are distinguished classes for the pupal and the larval antenna. Both are linked with an "isPartOf" relation to the corresponding developmental stages and share the same generic superclass "antenna". The annotations are linked to the classes connected to a developmental stage instead of being linked to generic ones. The most detailed level in the ontology is chosen for the annotation, i.e., flagellum is used if the phenotype affects only the flagellum and not the whole antenna. For the search interface the requirements are different. A typical input is the developmental stage and a generic

and rather abstract morphological structure, e.g., antenna instead of flagellum. To fulfill the demands and provide a general access to the *Tribolium* ontology the OBA service is embedded into the search interface and a server plugin with specific semantic functions has been implemented.

Upon startup the OBA service scans the ontology for concrete classes (these connected to a developmental stage) and generic classes, respectively. The concrete classes do not necessarily have a direct relation to a developmental stage, the path to the stage may be a collection of "is_a" and "isPartOf" links. The generated list of generic classes is used as a suggestion list for the user while typing into the search form. When the user has chosen a developmental stage and an anatomical structure, the OBA service selects all concrete classes downstream of the selected structures and connected to the appropriate stage. Because "isPartOf" is used in the *Tribolium* ontology to describe meronomic relation, the inverse "hasPart" relation is generated on the fly. The list of ontology classes is used as input for the search in the database of the iBeetle project. As add-on on the result page a tree with the subsections of the ontology that were used for the search is displayed. **Figure 5** shows a screenshot of this ontology tree. The semantic search started with the search term "head" and added all ontology classes representing head and its parts.

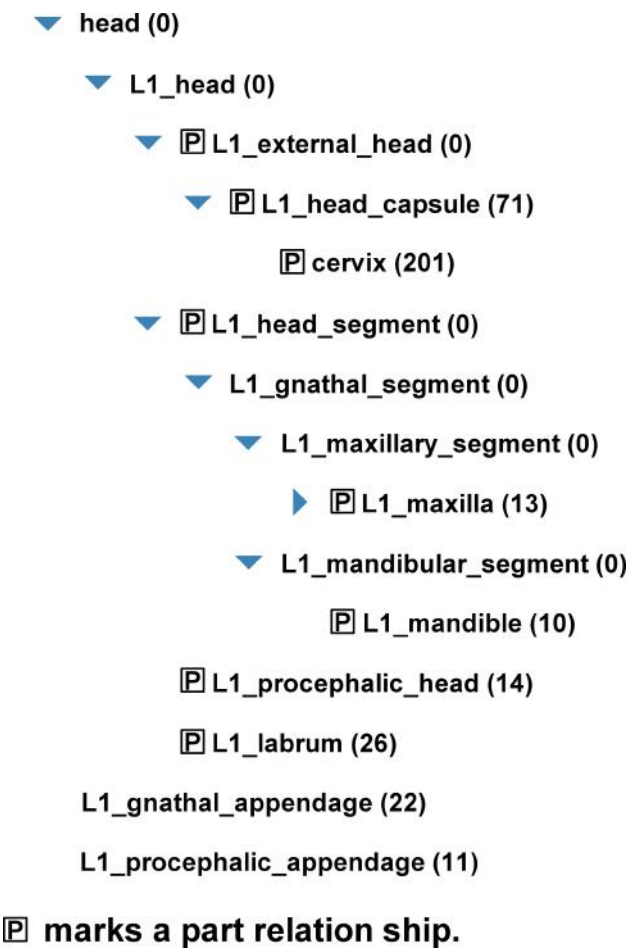


FIGURE 5 | Ontology tree from the result page of a search in the iBeetle database. The tree shows the classes of the *Tribolium* ontology downstream of the searched structure and linked to the queried developmental stage. In this example, the user has selected "head" as anatomical structure and "larva" as stage. All ontology classes shown in the tree where used for a search in the iBeetle database. The numbers in parentheses indicate the number of hits linked to this node.

PROJECT: EndoNet

For the upcoming new web interface for EndoNet, an information resource of the human endocrine system (Dönitz et al., 2008), a semantic search, similar to the search function described above is used. As ontological data source the anatomical ontology Cytomer is utilized. In this case the focus is not on developmental stages but on grouping the annotated cells and tissues at the level of organs in order to generate a survey map of general pathways. To limit the search result to anatomical structures used in EndoNet a predefined list containing the anatomical structures used in EndoNet is stored on the OBA server.

PROJECT: OntoScope

Another type of application using the OBA service is the ontology viewer OntoScope[9]. OntoScope visualizes ontologies as a graph extending the common tree like view of ontologies. The representation as a graph enables the user to explore ontologies along arbitrary relations. OntoScope uses from the OBA service the object graph and the access to the ontologies without any knowledge about the format or semantics of the ontology. OBA functions are used in the background, so that for example the nodes of the Cytomer ontology can be displayed in a color code according to the physiological system. **Figure 6** shows a screenshot of OntoScope with several nodes and relations.

Table 2 summarizes the OBA functions used in the projects. The plugin containing the function is named and a short description is given.

INSTALLATION AND EXTENSION OF OBA

For the use of OBA in a new application the Java client has to be downloaded and added to the class path of the application. After the initialization of the connector, all OBA functions are accessible as Java methods through the connector. The OBA functions will return single ontology classes or lists of them. These ontology classes are mapped to Java objects by the connector and returned by the Java methods of the connector. The Java objects provide functions to access the annotations and neighboring classes of the represented ontology class. If necessary missing information is queried internally from the OBA server. The application developer does not have to be concerned about the retrieval of neighboring classes.

If a required ontology is not available on the public OBA server, it can be downloaded and started locally. After the extraction of the zip file default directories for ontologies, plugins, and the storage area are available. New ontologies can be copied to the ontology directory together with a short property file. The property file defines under which name the ontology will be available from the OBA server and which annotation fields should be indexed for the

[9]http://www.bioinf.med.uni-goettingen.de/projects/ontoscope/

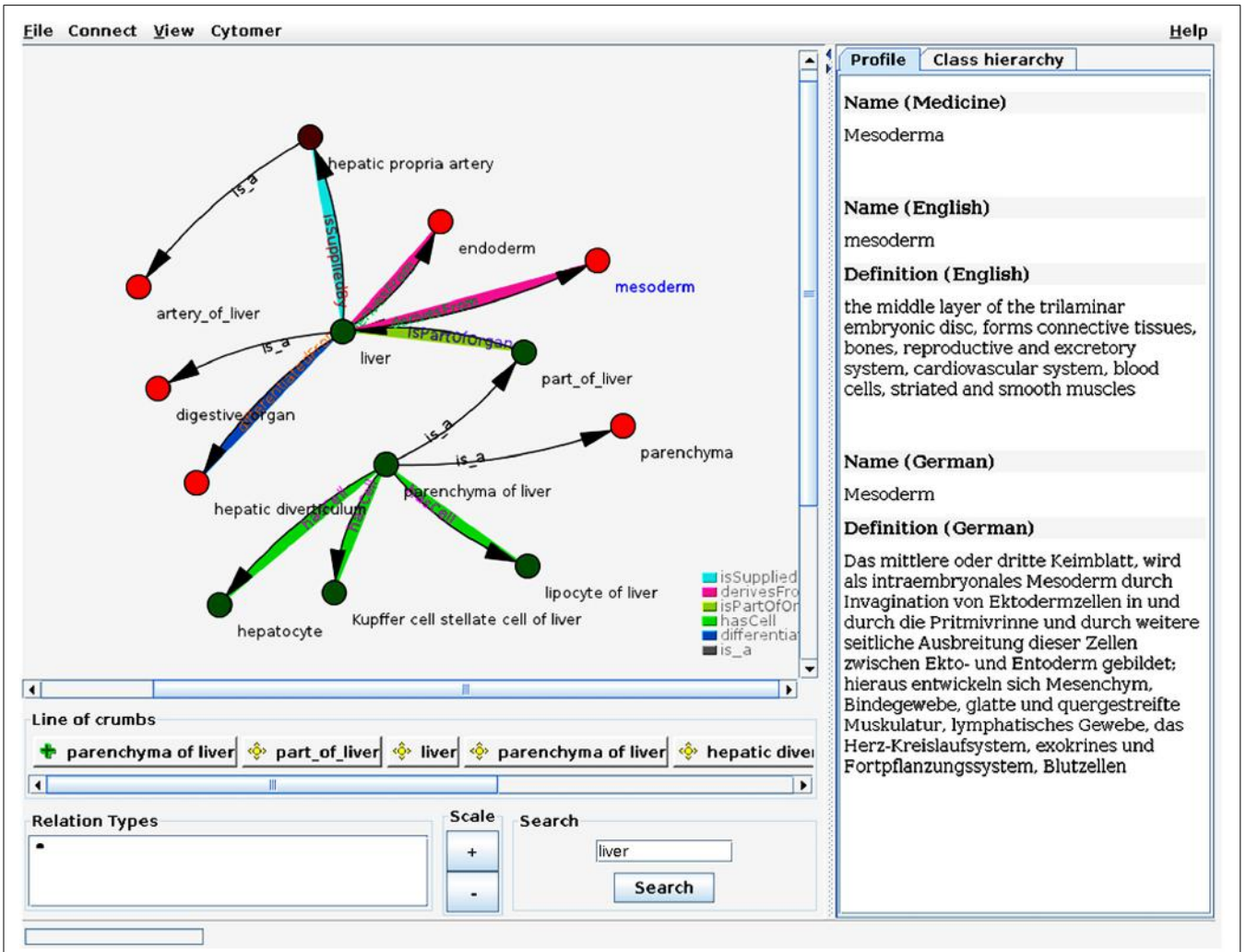


FIGURE 6 | Screenshot of OntoScope. The ontology viewer OntoScope uses the object graph from OBA to let the user browse the graph. The ontology classes are represented by the nodes of the graph. Each relation type has an own color and are displayed as edges connecting the nodes. The color of the nodes indicates the physiological system the class belongs to and is retrieved from the OBA service. On the tabs of the right side additional information of the selected ontology class is displayed, or the class hierarchy of the classes in the graph (tab is hidden in the screenshot).

search function. The property file can be copied from the provided examples and is described in the manual.

SUMMARY

The OBA service is available online at http://oba.sybig.de. Upon pointing a web browser to this URL an overview is given as a list of loaded ontologies as well as the available plugins and the OBA functions implemented by them. The object graph of the ontologies can be browsed by following the links of the HTML representation of the ontology classes. The syntax to access the OBA functions is described in the manual available at the home page of the project: http://www.bioinf.med.uni-goettingen.de/projects/oba. Located on the home page of the project is the Java connector as well as all sources and jar files for the server and currently available plugins. The Cytomer connector contains a test client, which is executed when the client is run on the command line. This client calls some functions on the server and prints the results to the console in order to validate the OBA service's function. The client's sources can serve as a template for a usage of OBA in a custom application.

To give the user a first impression of the function of the OBA service, a web demo is available at http://webdemo.oba.sybig.de/ implementing some of the provided functions for manual tests. For each step the example source code is noted, which is needed to implement the corresponding step in a custom application.

DISCUSSION

Ontologies are powerful and also complex tools. This is especially true for the OWL format. Parsers like the Jena-API (Jena – A Semantic Web Framework for Java[10]) or the OWL-API (Horridge and Bechhofer, 2011), take care of parsing ontologies but do not intend to hide the semantics of ontologies. The same is true for OBO ontologies, although they have a more finite

[10]http://jena.sourceforge.net

Table 2 | Overview of the OBA functions used in the projects.

Project	Used OBA function	Plugin	Functionality
iBeetle	concreteClasses	Tribolium	Returns all classes linked to a developmental stage. The annotated phenotypes are linked to these classes
	genericClasses	Tribolum	Returns all classes not related to a developmental stage, used for the auto-complete function of the search interface
	findInGeneric	Tribolium	Searches in the labels and synonyms of generic classes and an additional previous generated list for classes matching the search string. Used for the auto-complete function in the search interface
	concreteForDevStage	Tribolium	Returns the class downstream of the given generic class and linked to the given developmental stage. Used to map the user query to the annotations stored in the database.
EndoNet	findUpStreamInSet findDownStreamInSet	Cytomer	Used to find entities from EndoNet related to the search term
OntoScope	physiologicalSystemOf	Cytomer	Returns all physiological systems of an ontology class, used for coloring in the graph
	searchCls	built-in	Searches ontology classes matching a text pattern in the class name or annotation field

The table summarizes the use of OBA in the projects listed in the first column. The second and the third column denominate the OBA function name and the plugin containing the function. The last column describes the functionality of the OBA service the projects benefits from.

structure. If a developer plans to include information deduced from ontologies in an application, a time for training is needed to learn the semantics of ontologies and the framework's design. The basic tutorial of the OWL-API already consists of over 100 slides and deals with a semantic most computational biologists are unfamiliar with. The OBA service maps the relevant parts of ontologies to the world of object-oriented programming and provides semantic functions. The usage of the OBA service does not call for intensive training time to work with different topics and programming paradigms. The simplification to an object graph is oblivious to advanced features of OWL like cardinalities or different OWL dialects. If such a full access is needed, it can be achieved with the very good ontology APIs, i.e., Jena-API or OWL-API, with the query language SPARQL or Protege for interactive work. However, the OBA service can load and process any ontology in the OBO or OWL format, giving access to their fundamental information to developers who otherwise would probably not use ontologies.

Portals like OntoCAT (Adamusiak et al., 2011), the OLS (Côté et al., 2008), or the NCBO BioPortal (Noy et al., 2009) aim to provide access to huge collections of ontologies in a standardized manner. This is the preferred way if the unique definitions of terms in ontologies take precedence over the complex relations. Like the OBA service, OntoCAT and the NCBO ontology portal allow the user to access ontologies using the REST-protocol. OntoCAT also provides basic clients for different programming languages. In addition to the functions of the OntoCAT client, the Java objects of the OBA service provide the required functions to access the super- and subclasses as well as classes which are linked by relations. Together with the proxy function, the basis of the new feature in the OBA service is to map ontology classes to an object graph, traversable by Java methods. The required network communication with the service is encapsulated by the OBA client and transparent to the user. The feature to grant access to the neighbors of an object, representing an ontology class, by Java methods is beyond the function provided by the clients of the existing ontology portals. Together with the proxy function of the OBA client the developer is now enabled to access ontology classes and traverse the graph using only Java methods. Network access and parsing of the ontology is transparent.

One intention of the OBA service is to relieve the user from ontology specific demands by encapsulating the logic in a service. With the OBA functions the developer benefits from the rich information of a specific ontology encoded in the relations without the detailed knowledge about these semantics. The goal of the OBA service is not primarily to provide network access to ontologies, but to add additional functions to help a developer to solve a subtask of an application based on information available in ontologies without being familiar with ontologies, APIs, or query languages to process them.

The OBA service's concept of semantic functions is distinct from the goal of ontology portals like OBO-Foundry (Smith et al., 2007), NCBI, or OntoCAT. The portals focus on accessing as many ontologies as possible. This approach is very well suited for an ontology overarching search and access. The OBA service provides access to a set of specific ontologies with matching semantic functions. If a plugin with the required semantic function is already available the developer saves time for training and programming. Even if the required function is not available, the developer benefits from the framework of the OBA service and the advantages of the client described above. The OBA framework and the open architecture minimize the effort of extending the service to fit the requirements of a specific project. A new plugin relays on the existing functions to access the ontology, marshal the objects for the network transfer as well as the proxy functionality of the client. A new plugin only has to implement knowledge about a custom ontology or the logic to solve a new question. Due to the provided framework the already supplied plugins are very small and easy to implement. The developer of a new plugin needs to be familiar

with the curation guideline of the used ontologies. Further expertise about ontologies, like the different formats and ontology internals like Frames, Description Logic are not required.

Under the umbrella of the OBO-Foundry a collection of tools handling ontologies has evolved. There is a number of tools supporting the annotation process or focusing on statistical analysis of data based on ontologies, examples are the tool DAVID (Huang et al., 2009) and tools for the gene set enrichment analysis (GSEA) method (Subramanian et al., 2005). Like the functions of the OBA service, these tools make intensive use of the GO or other ontologies. The advantage of the OBA service is that it is easily extendible. The server can load plugins for any ontology. The service is designed to be embedded into applications and workflows to minimize interaction with external tools.

The design of the OBA service has several advantages. A public server is the central contact point and serves a growing collection of publicly available ontologies and plugins. Developers and maintainers of an ontology are welcome to submit new plugins, which enables the scientific community to profit. Alternatively, the server can be downloaded and run locally if the required ontology is not available in the public repositories, or if the developed plugin is not to be published.

The new features of OBA are the seamless mapping of ontologies to a connected object graph for object-oriented programming and the implementation of the OBA functions.

The server side plugins can make intensive use of the ontologies loaded by the server and return the computed results back to the client. The round-trips between client and server are reduced to a minimum and the logic is encapsulated in a reusable plugin. This new features enables computational biologists to use the basic information from ontologies in their applications, who would otherwise avoid ontologies.

ACKNOWLEDGMENTS

This work was supported by the Seventh Framework Program of the EU-funded "LipidomicNet" (grant no. 202272). We also acknowledge the support by Deutsche Forschungsgemeinschaft and Open Access Publication Funds of Goettingen University. The authors wish to acknowledge the help of Pia Franziska Kohlbecker with linguistic corrections.

REFERENCES

Adamusiak, T., Burdett, T., Kurbatova, N., van der Velde, K. J., Abeygunawardena, N., Antonakaki, D., et al. (2011). OntoCAT – simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics* 12, 218. doi: 10.1186/1471-2105-12-218

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Sci. Am.* 284, 34–43.

Bucher, G., Scholten, J., and Klingler, M. (2002)). Parental RNAi in *Tribolium* (Coleoptera). *Curr. Biol.* 12, R85–R86.

Burger, A., Davidson, D., and Baldock, R. (eds). (2008). *Anatomy Ontologies for Bioinformatics.* New York: Springer.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288–289.

Côté, R. G., Jones, P., Martens, L., Apweiler, R., and Hermjakob, H. (2008). The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.* 36, W372–W376.

Day-Richter, J., Harris, M. A., Haendel, M., Gene Ontology OBO-Edit Working Group, and Lewis, S. (2007). OBO-Edit – an ontology editor for biologists. *Bioinformatics* 23, 2198–2200.

Dönitz, J., Goemann, B., Lizé, M., Michael, H., Sasse, N., Wingender, E., et al. (2008). EndoNet: an information resource about regulatory networks of cell-to-cell communication. *Nucleic Acids Res.* 36, D689–D694.

Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures.* Ph.D. thesis, University of California, Irvine.

Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., et al. (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* 27, 318–322.

Horridge, M., and Bechhofer, S. (2011). The OWL API: a Java API for OWL ontologies. *Semant. Web* 2, 11–21.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* 2011, bar030.

Kurbatova, N., Adamusiak, T., Kurnosov, P., Swertz, M. A., and Kapushesky, M. (2011). ontoCAT: an R package for ontology traversal and search. *Bioinformatics* 27, 2468–2470.

Lacy, L. W. (2005). *OWL: Representing Information Using the Web Ontology Language.* Victoria: Trafford Publishing.

Michael, H., Chen, X., Fricke, E., Haubrock, M., Ricanek, R., and Wingender, E. (2005). Deriving an ontology for human gene expression sources from the CYTOMER database on human organs and cell types. *In Silico Biol.* 5, 61–66.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., et al. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 37, W170–W173.

Schröder, R., Beermann, A., Wittkopp, N., and Lutz, R. (2008). From development to biodiversity–Tribolium castaneum, an insect model organism for short germband development. *Dev. Genes Evol.* 218, 119–126.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Tomoyasu, Y., and Denell, R. E. (2004). Larval RNAi in *Tribolium* (Coleoptera) for analyzing adult development. *Dev. Genes Evol.* 214, 575–578.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* 28, 1248–1250.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* 7, e1000247. doi: 10.1371/journal.pbio.1000247

Winston, M. E., Chaffin, R., and Herrmann, D. (1987). A taxonomy of part-whole relations. *Cogn. Sci.* 11, 417–444.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 March 2012; accepted: 14 September 2012; published online: 05 October 2012.

Citation: Dönitz J and Wingender E (2012) The ontology-based answers (OBA) service: a connector for embedded usage of ontologies in applications. Front. Gene. 3:197. doi: 10.3389/fgene.2012.00197

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.

frontiers in GENETICS

OPINION ARTICLE
published: 03 September 2013
doi: 10.3389/fgene.2013.00171

Social networks for eHealth solutions on cloud

Briti Deb* and Satish N. Srirama

Mobile Cloud Laboratory, Institute of Computer Science, University of Tartu, Tartu, Estonia
**Correspondence: deb@ut.ee*

Edited by:
John Hancock, University of Cambridge, UK

Reviewed by:
Jenny Ure, Edinburgh University, UK

Keywords: cloud computing, eHealth, crowdsourcing, social networks, semantic web, participatory medicine, utility computing/on-demand computing

INTRODUCTION

Traditional experimentation based healthcare solutions are constrained by limited data that can confirm or refute the initial hypothesis. Big medical data in individual Electronic Health Records, labs, imaging systems, physician notes, medical correspondence and claims, provides a resource for extracting complementary information that can enhance the data available from traditional approaches based on experimentation. Datamining algorithms are being used to analyze data to get a more insightful understanding of human health, both preventive and clinical. But despite their sophistication, they are far from flawless. One way to solve the problem is crowdsourcing citizens connected in a social network, who can provide data, get it analyzed, and consume data for preventive health insights (Swan, 2009). Several challenges come along with it, for instance: performance, scalability, speed, storage, and power, which we believe could be addressed by cloud-enabled social networks for eHealth services. Such services could be composed of many other services, for instance, user authentication, email, payroll management, calendars, tele-consultation, e-Prescribing, e-Referral, e-Reimbursement, and alerting services, aiming to change the way big medical data in social networking web sites could be used making it actionable to save lives.

This paper aims to explore the opportunities and challenges for realization of cloud-enabled social networks for eHealth solutions, by examining efforts already underway, and recommending solutions to improve it. We discuss a three-tier ecosystem to advance this key field leveraging the Cloud computing technologies. In Tier-1 is "*Build Sustainable eHealth System*" to create a foundation that facilitates secure creation, storage, exchange, and analysis of data between actors. In Tier-2 is "*Crowdsourced Social Networks for eHealth Services*" to utilize the power of crowdsourcing. In Tier-3 is "*Increasing Access to eHealth*" to minimize risk and improve patient outcome. Failure to address these issues is believed to result in inefficient use of big medical data toward preventive healthcare.

THE THREE-TIER eHEALTH ECOSYSTEM ON CLOUD

TIER 1: BUILD SUSTAINABLE eHEALTH SYSTEM

Semantic interoperability

As healthcare institutes do not strictly conform to a single commonly agreed vocabulary/standard, integrating biomedical data using domain ontologies is far from perfect (Della Valle et al., 2005; Ulieru et al., 2006). We believe that such diverse data in terms of volume, variety, and velocity, can be attempted to be semantically integrated, shared, reused, and made accessible, by using a top-level ontology for integrating domain ontologies, semantic web standards such as RDF for describing information, SPARQL as an RDF query language, and OWL to represent knowledge.

Compliance/accountability

Specific compliance/accountability requirements can be enforced by laws and regulations on organizations that collect, generate or store medical data, thereby dictating a wide array of data related policies such as, retention time, deletion process, recovery plans, and sharing policy. Laws such as the Health Insurance Portability and Accountability Act (HIPAA) in the US are already in force and complied with by organizations like *PatientsLikeMe.com.* The Federal Risk and Authorization Management Program (FedRAMP) is another law in the US enacted to assess and authorize cloud products and services. The dispersed geographic location of cloud providers such as *Amazon.com* opens the possibility of breach of compliance, which could be addressed by *Portable Consent*, and Institutional Review Board could be enacted to monitor, approve, or prevent the use of medical data on the cloud.

Security and privacy

Hosting data in the cloud poses privacy concerns because the service provider may access, accidentally or deliberately alter, or even delete information. Methods to obfuscate individual identity attributes such as Zero-knowledge Technology or Privacy Enhancing Technologies are currently not used in a pervasive manner (Bertino et al., 2009) due to lack of granularity in the Access Control List, creating privacy risks. To mitigate some of the security risks such as sensitive data access, data segregation, bug exploitation, recovery, accountability, and activity by malicious insiders, solutions are being researched such as cryptography, public key infrastructure (PKI), standardisation of APIs, and virtual machine security.

Legislative influence

As the Cloud poses a challenge on "possession," "custody," and "ownership" of data, Terms of Service (TOS) agreements become vital to clarify the different rights to be assigned to different roles. The TOS must also specify procedures to follow in the event of an end of provider-customer relationship, a merger of one provider with another, bankruptcy, and insolvency. An open challenge is how to ascertain legal jurisdiction if disputes arise for geographically dispersed data. Patient Advocacy Groups could play a role in influencing advisory panels toward adopting better laws to protect providers and consumers.

Revenue/financial model

Crowdsourced eHealth social networks are mostly free of subscription fees, advertising, banner ads or popups. Sale of anonymized data, clinical trial awareness programs, and market research surveys constitute a major part of revenue. In future, revenue model could increasingly include health insurers, such as the already implemented *Health Savings Account* in US.

Reputation/credibility, quality control, and transparency

The success of safety-critical systems depends largely on the reputation/credibility they enjoy in market. Several non-technical challenges arises from the change in the IT department's role from provider to consultant (Khajeh-Hosseini et al., 2010), resulting in an increased risk to customer satisfaction, job quality, and job satisfaction, tensions between the expectations of different groups, questioning the long term organizational impact of Cloud migration on reliability, scalability, and cost effectiveness.

TIER 2: CROWDSOURCED SOCIAL NETWORKS FOR eHEALTH SERVICES

Personalized preventive health maintenance comes against the backdrop of several challenges such as difficulty in understanding the causations of complex diseases due to an incomplete understanding of the complexities of biology, the high cost of healthcare, an aging population, and a physician shortage. One solution is to use social networks as a platform to facilitate the participation of millions of users in the crowd to realize the 4P's of medicine—preventive, personalized, predictive, and participatory. Several eHealth social networks have appeared, namely, *patientslikeme.com, hellohealth.com, medhelppc.org, curetogether.com, dailystrength.org, FacetoFaceHealth.com, 23andMe.com, Genomera.com, QuantifiedSelf.com, DIYgenomics.org*, providing a platform for people in the crowd to compare their conditions with other individuals, and identifying areas for further scientific research on their own before clinical symptoms appear. Studies have shown typical challenges for a crowdsourced system (Doan et al., 2011) such as (a) recruitment, retention, and evaluation of users, (b) merging/combining contribution of users, (c) managing quality of contribution of users, (d) managing query semantics, query execution, and query optimization, and (e) improving user interfaces.

In addition to identifying potential preclinical symptoms, datamining algorithms can be applied to the discussion forums provided by the eHealth social networks to identify epidemiological patterns such as (i) patient behavior in response to a safety event, (ii) efficacy and side-effects of drugs that have not shown up in trials, thereby helping to reduce time spent in clinical trial, (iii) monitoring and participating in real-world natural experiments, (iv) anonymously sharing treatment, symptom, progression and outcome data.

However, performance and adaptability of eHealth social networks face challenge due to complexities in big data handling, such as variety, velocity, volume, distribution, synchronization, fault recovery, etc. To address the challenge of distributing data and computation loads over multiple processing units, largely three main directions have being studied: (a) parallel computing frameworks such as MapReduce, Iterative MapReduce, and Bulk Synchronous Parallel (BSP), (b) Graphics Processing Units, and (c) Message Passing Interfaces.

In the MapReduce model, parallelism is achieved by executing Map and Reduce tasks concurrently. To achieve fault tolerance, data is replicated and failed tasks are re-executed. The efficiency and scalability of algorithms on the Cloud can be affected by the characteristics of an algorithm, necessitating a classification for algorithms (Srirama et al., 2012). As the MapReduce model is most suitable for embarrassingly parallel tasks, i.e., parallel tasks having little or no dependency between them, serious issues arise when working with graph problems in social networks due to factors such as (a) long "*start up*" and "*clean up*" times, (b) no way to keep important data in memory between MapReduce job executions, and (c) reading of all data from file system (HDFS) after each iteration and writing back there at the end. Three main directions are currently being pursued to address the challenges of graph processing in parallel environment: (i) restructuring algorithms for the non-iterative MapReduce version, (ii) restructuring non-iterative MapReduce algorithms into iterative MapReduce versions using alternative MapReduce frameworks (Twister, HaLoop, Spark), giving up advantages of the MapReduce model such as Fault tolerance and running multiple concurrent reduce tasks, and (iii) alternative distributed computing models such as BSP (Pregel, Hama, Giraph).

TIER 3: INCREASING ACCESS TO eHEALTH

Several challenges limit access to eHealth. One such is the workflow challenge, arising for several reasons such as the inefficiency of current processes and the dependency on paper to store data. It is envisioned that in future, a physician would enter patient data in an electronic scheduling system on the Cloud, which would be processed by some workflow to automatically determine the most appropriate test, and the patient directly notified of the possible options.

Semantically integrating diverse patients medical records, census data, and environmental samplings, and managing scalability and load balancing, are some of other major challenges while analyzing big data. One approach to addressing these is the use of virtualization technology, which allows applications to be easily migrated from one physical server to another, resulting in improved reliability, scalability, business continuity, load balancing, hardware maintenance, disaster recovery, and better utilization of processors and memory.

Yet another challenge to increasing access to healthcare is providing ubiquitous healthcare monitoring. Traditionally, patients were "treated" only in hospital/clinic, which is expected to change in future, as ubiquitous gadgets such as mobile phones are now being increasingly being used to track patients and keep them compliant. Mobile cloud computing is expected to arise as a prominent domain, seeking to bring the massive advantages of the Cloud to resource constrained smartphones, by following either the *delegation model* or *code offloading* model (Flores and Srirama, 2013). In the *delegation model*, a mobile phone consumes services from multiple clouds by following their Web API, whereas, in the *code offloading* model,

a mobile application is partitioned and analyzed so that the most computationally expensive operations at code level can be identified and offloaded to the Cloud for remote processing.

DISCUSSION AND CONCLUSIONS

In this paper, we briefly analyzed the opportunities and challenges for realization of cloud-enabled social networks for eHealth solutions, and proposed a three-tier ecosystem to improve it. Four main actors can be identified: service providers (genomic counselors, biomedical researchers), remedy providers (eHealth social networks providing computing and storage), health professionals, and data provider/consumers. The challenges can be summarized into two main groups. First, technical challenges such as resource exhaustion attributed to the ever increasing demand of the Cloud resources, data transfer bottlenecks attributed to the limited network bandwidth, unpredictability of Cloud performance attributed to the inability of Cloud consumers to govern the virtual architecture owned by Cloud providers, data lock-in attributed to the discontinuity of Cloud-based eHealth services, compounded by the problem of semantic interoperability when migrating the data to another Cloud, and limitations of the non-iterative MapReduce model, particularly in scalable graph processing. Second, non-technical challenges arising from the change in the IT department's role from provider to consultant, affecting customer satisfaction and overall service quality, calling for stringent quality control and transparency measures. To address these issues, we proposed a three-tier eHealth ecosystem. In future, we propose to: (i) investigate the use of Parallel R packages to leverage multi-processor systems to speed computations with big data by explicit parallelism, implicit parallelism, and implementing map-reduce for Hadoop; (ii) develop novel algorithms for parallel classification and parallel search; and (iii) develop a novel framework for semantic integration of biological data in social networks leveraging the Cloud. We believe that a combined strategy consisting of semantic, algorithmic, and computational approaches would be useful to solve many problems in eHealth social networks on the Cloud. Biological research would benefit as researchers would be able to analyze massive amounts of complex data much more quickly, and generate hypotheses faster. Finally, the authors believe that research in that direction could enhance the scale and scope of experiments that are possible, resulting in an exponential growth in knowledge, similar to the exponential growth in data that we see today.

ACKNOWLEDGMENTS

This research is supported by the European Regional Development Fund through the EXCS, Estonian Science Foundation grant ETF9287, Target Funding theme SF0180008s12 and European Social Fund for Doctoral Studies and Internationalization Programme DoRa.

REFERENCES

Bertino, E., Paci, F., Ferrini, R., and Shang, N. (2009). Privacy-preserving Digital Identity Management for Cloud Computing. *IEEE Data Eng. Bull.* 32, 21–27.

Della Valle, E., Cerizza, D., Bicer, V., Kabak, Y., Laleci, G., Lausen, H. et al. (2005). "The need for semantic web service in the eHealth," in *W3C workshop on Frameworks for Semantics in Web Services*, (Innsbruck).

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *CACM* 54, 86–96.

Flores, H., and Srirama, S. N. (2013). "Adaptive Code Offloading for Mobile Cloud Applications: Exploiting Fuzzy Sets and Evidence-based Learning," in *The Fourth ACM Workshop on Mobile Cloud Computing and Services (MCS 2013), at The 11th International Conference on Mobile Systems, Applications and Services (MobiSys 2013)*, (Taipei, Taiwan: ACM), 9–16.

Khajeh-Hosseini, A., Greenwood, D., and Sommerville, I. (2010). "Cloud migration: a case study of migrating an enterprise it system to iaas," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on. IEEE* (Miami, Florida).

Srirama, S. N., Jakovits, P., and Vainikko, E. (2012). Adapting scientific computing problems to clouds using MapReduce. *Future Gen. Comput. Syst.* 28, 184–192. doi: 10.1016/j.future.2011.05.025

Swan, M. (2009). Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *Int. J. Environ. Res. Public Health* 6, 492–525. doi: 10.3390/ijerph6020492

Ulieru, M., Hadzic, M., and Chang, E. (2006). Soft computing agents for e-Health applied to the research and control of unknown diseases. *Inf. Sci.* 176, 1190–1214.

Received: 17 June 2013; accepted: 15 August 2013; published online: 03 September 2013.

*Citation: Deb B and Srirama SN (2013) Social networks for eHealth solutions on cloud. Front. Genet. **4**:171. doi: 10.3389/fgene.2013.00171*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.