DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bas057, doi:10.1093/database/bas057

Original Article

Matching arthropod anatomy ontologies to the Hymenoptera Anatomy Ontology: results from a manual alignment

Matthew A. Bertone[1,*], István Mikó[1,2], Matthew J. Yoder[1,3], Katja C. Seltmann[1,4], James P. Balhoff[5,6] and Andrew R. Deans[1,2]

[1]Department of Entomology, North Carolina State University, Campus Box 7613, Raleigh, NC 27695-7613, [2]Department of Entomology, Pennsylvania State University, 501 ASI Building, University Park, PA 16802, [3]Illinois Natural History Survey, 1816 South Oak Street, Champaign, IL 61820, [4]Department of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th St., New York, NY 10024-5192, [5]National Evolutionary Synthesis Center, Durham, NC 27705 and [6]Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, USA

***Corresponding author:** Tel: 919-210-9857; Fax: 919-515-7746; Email: matthew.bertone@gmail.com

Citation details: Matthew A. Bertone, István Mikó, Matthew J. Yoder, *et al.* Matching arthropod anatomy ontologies to the Hymenoptera Anatomy Ontology: results from a manual alignment. *Database* (2012) Vol. 2012: article ID bas057; doi:10.1093/database/bas057.

Submitted 3 August 2012; Revised 31 October 2012; Accepted 27 November 2012

Matching is an important step for increasing interoperability between heterogeneous ontologies. Here, we present alignments we produced as domain experts, using a manual mapping process, between the Hymenoptera Anatomy Ontology and other existing arthropod anatomy ontologies (representing spiders, ticks, mosquitoes and *Drosophila melanogaster*). The resulting alignments contain from 43 to 368 mappings (correspondences), all derived from domain-expert input. Despite the many pairwise correspondences, only 11 correspondences were found in common between all ontologies, suggesting either major intrinsic differences between each ontology or gaps in representing each group's anatomy. Furthermore, we compare our findings with putative correspondences from Bioportal (derived from LOOM software) and summarize the results in a total evidence alignment. We briefly discuss characteristics of the ontologies and issues with the matching process.

Database URL: http://purl.obolibrary.org/obo/hao/2012-07-18/arthropod-mappings.obo

Introduction

Representing information about a domain of interest as an ontology is an increasingly important way to formalize concepts and aid computer reasoning of real-world systems. Although ontologies have been created for many domains, bio-medicine contains some of the most complex examples, reflecting the intricacy of nature. Within this domain, several ontologies have been developed to model the anatomy (morphology) of arthropods (Metazoa: Ecdysozoa: Arthropoda), the largest and most diverse group of organisms on Earth. Five arthropod taxa have representative anatomy ontologies on the Open Biological and Biomedical Ontologies (OBO) Foundry (1): spiders [Arachnida: Araneae; SPD; (2)], ticks [Arachnida: Ixodida; TADS; (3)], mosquitoes [Insecta: Diptera: Culicidae; TGMA; (3)], *Drosophila melanogaster* [Insecta: Diptera: Drosophilidae; FBbt; (4)] and wasps and their relatives [Insecta: Hymenoptera; HAO; (5)]. These ontologies range in size from 552 (SPD) to 6884 (FBbt) valid classes (at the time of analysis; Table 1) and differ in general content, structure and granularity. The disparity in size and scope of these ontologies is primarily due to their varied purposes, organization and intended audience. For example, the ontology we created and curate, HAO, was developed to aid in standardizing the meaning of anatomical concepts used by taxonomists to describe the insect order Hymenoptera, while also providing a way to reason across large sets of

Database, Vol. 2013, Article ID bas057, doi:10.1093/database/bas057

Table 1. General statistics of the ontologies examined in this article

Subject (Ontology[a])	# Valid classes	# Obsolete classes	Proportion of classes with definitions (%) (total with definitions)	# Species covered currently (potential coverage)	Version date
Hymenoptera (HAO)	1786	64	100 (1786)	~150 000 (~1 million)	24:01:2011 09:40
Spiders (SPD)	552	25	73 (404)	~40 000 (~150 000)	17:03:2010 06:57
Ticks (TADS)	628	0	99 (627)	~900	18:11:2007 11:42
Mosquitoes (TGMA)	1861	0	100 (1861)	~3500 (~4500)	04:02:2009 10:45
Drosophila melanogaster (FBbt)	6884	162	47 (3239)[b]	1	24:11:2010 15:26

[a]Full ontology names from the OBO Foundry [http://www.obofoundry.org/; (1)] are as follows: HAO, Hymenoptera Anatomy Ontology; SPD, Spider Ontology; TADS, Tick gross anatomy; TGMA, Mosquito gross anatomy; FBbt, *Drosophila* gross anatomy. [b]275 of these definitions are represented only by '.'; the percentage of worded definitions is 43% (2964).

descriptive text to extract information that is not apparent when looking at the data independently (5). The remaining arthropod ontologies have other stated purposes, including annotating vector genomes (TGMA and TADS) (3) and classifying images for phylogenetic characters (SPD) (2), to name a few.

While their stated purposes are different, it follows that information within each ontology (and the external data that are connected to each) could benefit other ontologies, and probably should in some way. To overcome the heterogeneity among these ontologies, therefore, requires linking their information in a way that increases interoperability; this is usually accomplished through ontology matching and results in an alignment (6). Strategies for ontology matching have mainly focused on improving algorithms for automation of the process, to avoid time-consuming manual methods and the need for domain expert input. However, end users are still in need of authoritatively vetted alignments to make real-world queries and discoveries, and automation is not without its drawbacks and limitations (7). Thus far only a few alignments have been produced among organismal anatomy ontologies, such as those between mice and humans (8, 9) and multiple anatomy ontologies [Uberon (10, 11)].

As bioinformatics tools, ontologies are expected to aid in some level of discovery that cannot be achieved by looking at individual elements alone (12). Therefore, we expect queries that employ the logical reasoning built into ontologies to become more efficient, powerful and easier to implement (broadening user base). For example, one of the questions we as domain experts are interested in is the underlying genetics of various phenotypes exhibited by hymenopterans, an important query relevant to functional morphology, evolutionary developmental biology (evo-devo) and systematics. While there exist an abundance of genomic data from arthropod model organisms, forming meaningful, genetics-based hypotheses from the ontologies of these taxa is difficult because of their current state of relative insularity from each other. However, the premise exists that basic phenotypic data can be shared across taxa through an alignment of their anatomy ontologies. The resulting linkages facilitate the transfer of knowledge between domains.

Here, we present results from a domain expert-driven manual alignment of arthropod anatomy ontologies to the Hymenoptera Anatomy Ontology (HAO). Our aims were to (i) identify mappings (from here on referred to as correspondences) between the HAO and other arthropod ontologies and represent them as an alignment, (ii) compare the results of our manual approach with a currently available algorithmic dataset (LOOM mappings on Bioportal) and (iii) briefly discuss issues encountered while performing these manual alignments. We anticipate this to be a first step and expect the process to be repeated, allowing the results to be modified as the current ontologies grow and new anatomy ontologies are developed for other arthropod taxa.

Materials and methods

Ontologies were downloaded as OBO format files from the OBO Foundry (1); versions and general statistics of each are listed in Table 1. A manual alignment between HAO and the other arthropod ontologies was initiated by MAB and further refined by IM, both domain experts in arthropod anatomy (Diptera and Hymenoptera, respectively). Classes from each source ontology (SPD, TADS, TGMA or FBbt) were identified as matches to classes known in our target ontology (HAO), manually, using spreadsheets. Correspondences were based on lexical similarity (i.e. same name or label of the class, or of its synonyms when present) with additional evidence to avoid blindly matching homonyms (see 'Discussion' section), physical/structural similarity, evidence from definitions, evidence from figures in referenced texts and, sometimes, based on class relations such as subsumption or property restrictions. The structure of the ontologies was often modeled differently for similar classes, thus structure represented by relations was not generally an accurate arbiter for correspondences

(see 'Discussion' section). Other types of mappings, such as disjoint classes or more general classes were not described, as focus was limited to similar/congruent classes. Along those lines, homology [as defined in (13)] was a primary criterion for matching classes, but was not the only type of similarity used in our searches, as it is sometimes difficult to determine without direct observation and knowledge of the organisms' development. For example, the class for the hymenopteran basitarsus (the proximal tarsomere of each leg) was aligned with the spider class for metatarsus based on a similar position on the leg; this may not represent a homologous segment in both organisms. Literature examined for aligning classes included major works on the anatomy of the groups presented herein (14–18). Although 1:1 correspondences were most common and desirable, on several occasions other levels of cardinality (n:1, 1:m or n:m) were required (Figure 1), for example when multiple classes in one ontology were characterized as only one class in the other ontology. All alignments were translated into an OBO-format XREF alignment that is available at: http://purl.obolibrary.org/obo/hao/2012-07-18/arthropod-mappings.obo.

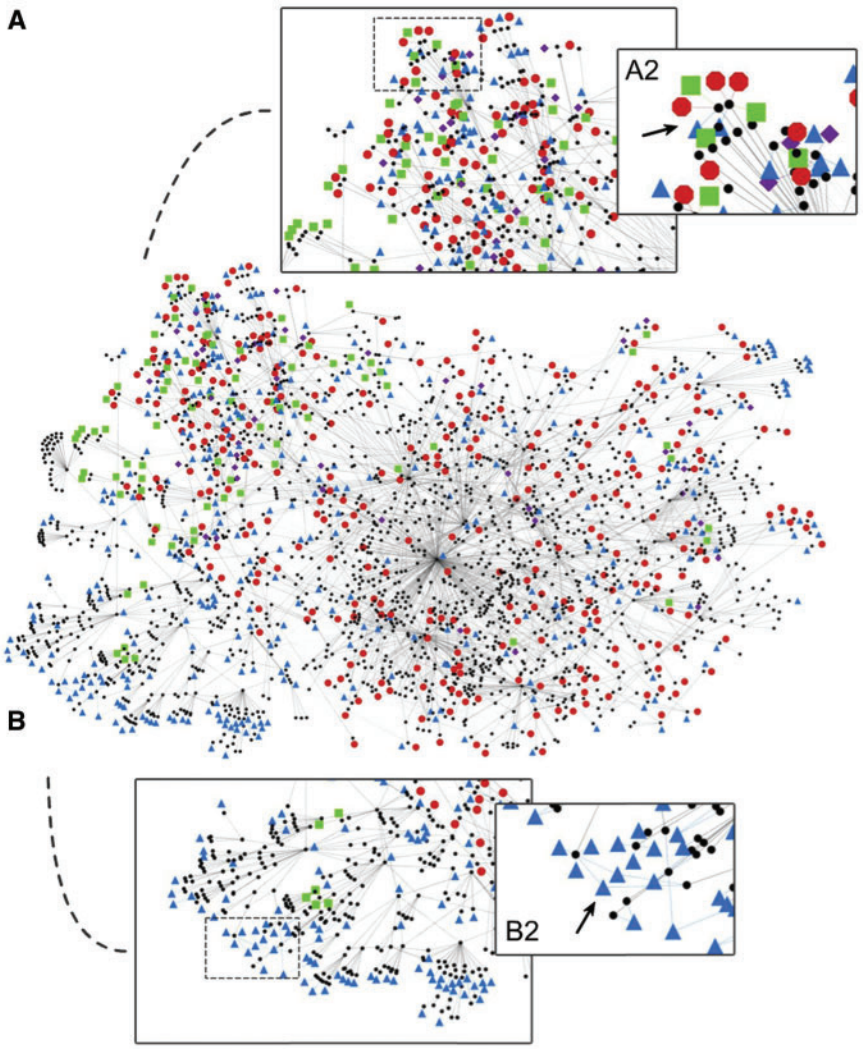

Figure 1. Cytoscape visualization showing the HAO full ontology network (black circle nodes and gray lines) with correspondences mapped from the other arthropod anatomy ontologies: SPD (purple diamonds); TADS (green squares); TGMA (red octagons); FBbt (blue triangles). Box A represents an area of general agreement between the ontologies, showing multiple correspondences from each ontology (largely consisting of CARO and many general body classes); further magnification (A2, represented by dashed box in A) reveals nodes with many correspondences from different ontologies (arrow). Box B represents an area with fewer correspondences, mainly from FBbt with some TADS (largely consisting of specific muscle classes not present in TGMA and SPD); further magnification (B2, represented by dashed box in B) reveals instances where one class from FBbt is aligned to multiple HAO classes (arrow; many to one relationship).

Database, Vol. 2013, Article ID bas057, doi:10.1093/database/bas057

To compare our manual alignment findings with automated/algorithmic results, we evaluated the mappings created using Lexical OWL Ontology Matcher or LOOM [http://www.bioontology.org/wiki/index.php/LOOM; (19)], publicly available on Bioportal (http://bioportal.bioontology.org/) and accessed on 15August 2011. Since mappings to HAO were only available from SPD, TGMA and FBbt (lacking for TADS), only those comparisons are presented here. Our evaluation consisted of gathering correspondences found by the manual process alone, LOOM alone or those common to both methods. We then noted whether the correspondences found only by LOOM were valid (overlooked during the manual process) or invalid (mismatches or other incorrect proposals). All results were quantified to compare the overall accuracy of the two major methods.

To facilitate exploration of the correspondences and other shared features of the arthropod ontologies, a small script library 'obo_parser' was developed. The code is available as a Ruby Gem (http://rubygems.org/), with source available at https://github.com/mjy. The library's core functionality is a set of tools for parsing the OBO file format (http://www.geneontology.org/GO.format.shtml). A set of utility methods are built on top of the parser and allow for conversion, for example, of ontology IDs to labels for tab-delimited columns of IDs. The utilities also include functionality that returns column-based reports of OBO labels and their relationships, suitable for seeding a Cytoscape-based visualization (http://www.cytoscape.org/). Cytoscape was subsequently used to visualize the HAO ontology structure with putative correspondences mapped from the other ontologies. All supporting data, including tables of correspondences, versions of the OBO files used and Cytoscape visualizations, have been deposited on Dryad (http://datadryad.org/).

Results

The following numbers of correspondences were found between each ontology and the HAO (Table 2 and Figure 1): 43 (SPD), 82 (TADS), 307 (TGMA) and 368 (FBbt). A list of all correspondences can be found in the alignment file: http://purl.obolibrary.org/obo/hao/2012-07-18/arthropod-mappings.obo. The relative proportion of correspondences to total classes ranged from ~2% to 21% (Table 2), meaning the general uniqueness of the ontologies relative to the HAO ranged from ~79% to 98%. Furthermore, though classes from the Common Anatomy Reference Ontology (20) were used as higher level, base classes, the portion of CARO itself used by each ontology ranged from ~8% (SPD) to 94% (TADS) and resulted in the variability of its contribution to the aligned correspondences (~5–55% of correspondences coming from CARO matches; Table 2). Finally, the intersection of the HAO, TADS and SPD resulted in 15 correspondences, while the intersection of the HAO, TGMA and FBbt had 151 correspondences (Appendix A–C). All five ontologies shared 11 correspondences (HAO class labels): anatomical entity (CARO), portion of organism substance (CARO), acellular anatomical structure (CARO), coxa, female genitalia, femur, leg, pretarsus, tarsal claw, tibia and trochanter.

The results from the automated method (LOOM; see 'Materials and Methods' section) differed from the manual alignment in both number of correspondences and degree of overlap (Figure 2 and Appendix D–F). The number of correspondences found by LOOM was as follows: 47 (SPD–HAO), 526 (TGMA–HAO) and 205 (FBbt–HAO). Furthermore, a comparison of the methods revealed these results (Figure 2): between the HAO and SPD, 34 correspondences were identified by both methods, 9 by manual alignment only and 13 by LOOM alone; between the HAO and TGMA, 152 correspondences were identified by both methods, 155 by manual alignment only and 374 by LOOM alone; between the HAO and FBbt, 132 correspondences were identified by both methods, 236 by manual alignment only and 73 by LOOM alone. Although it appears that in some cases the LOOM algorithm was more productive (see TGMA), an evaluation of its findings showed that many were mismatches (92% in the case of TGMA; Figure 2) as identified by domain experts. Thus, the actual number of valid improvements over those found by both methods were as follows (manual/algorithm): SPD—9/3; TGMA—155/7; FBbt—236/16. We also observed apparent algorithm errors from LOOM resulting in improper-recognition artifacts, including the use of obsolete classes (i.e. those classes that have been deprecated following the creation of newer, more accurate classes) from the HAO and the reuse of classes with *alt_id* fields (only yet identified from FBbt). LOOM also failed to recognize several valid correspondences (2, 4 and 18 in SPD, TGMA and FBbt, respectively) involving exact lexical matches that were validated during the manual alignment.

Each ontology defined its own set of relations, numbering from 1 (SPD and TADS) to 20 (FBbt) (excluding relations that are built into the OBO format, e.g. *is_a*; http://www.geneontology.org/GO.format.obo-1_4.shtml) (Table 3). Only one relation, *part_of*, was shared among the ontologies, either corresponding exactly to (FBbt) or inferred to be the same (SPD, TADS andTGMA) as the HAO.

Discussion

As expected, most correspondences were found between the two fly ontologies (TGMA and FBbt) and the wasp ontology (HAO), being that all three are closely related phylogenetically (Insecta: Holometabola) and, therefore, share a number of anatomical features. Conversely, the ontologies for spiders and ticks (Arachnida), more distantly related arthropods, had fewer correspondences with the

Table 2. Summary of correspondences found during the manual alignment process between source arthropod ontologies and the target ontology, the HAO

Source ontology	# Correspondences	Correspondences as % of valid classes (source/target)	# Correspondences from CARO aligned[a] (% of total)	# Correspondence direct superclass *is_a* matches[b] (yes/no)	# Correspondence direct superclass *part_of* matches[b] (yes/no)
Spiders (SPD)	43	7.8/2.4	4 (9.3)	14/10	5/13
Ticks (TADS)	82	13.1/4.6	45 (54.9)	49/14	12/19
Mosquitoes (TGMA)	307	16.5/17.2	30 (9.8)	79/84	85/252
Drosophila melanogaster (FBbt)	368	5.4/20.6	18 (4.9)	97/327	35/118

[a]Number of HAO Common Anatomy Reference Ontology (19) classes aligned (some putative CARO classes were not cited as belonging to CARO in all source ontologies). [b]Number of direct *is_a* or *part_of* superclasses that are (yes) or are not (no) additionally represented as matched correspondences (e.g. if A *is_a* C and B *is_a* D, then 'yes' if the correspondences A to B and C to D are present; if A and B correspond, but C and D do not, then 'no').

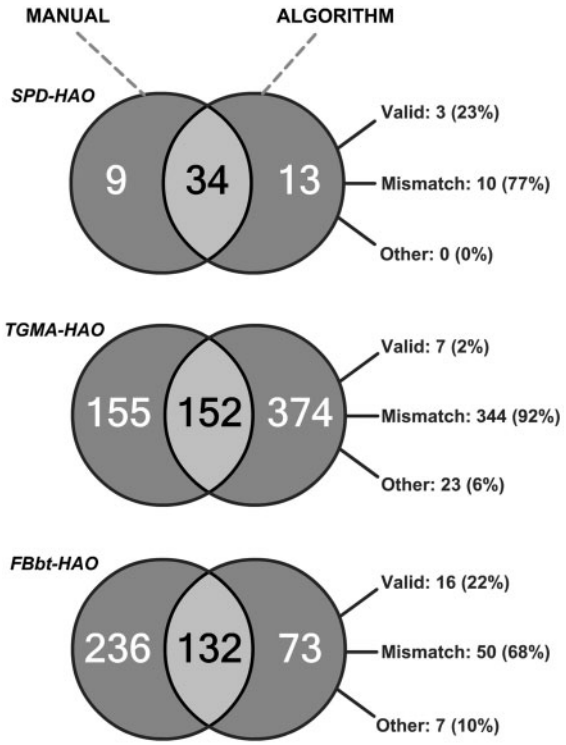

Figure 2. Comparison of the number of correspondences found through manual alignment alone, LOOM-based algorithm alignment alone (available from Bioportal) and using both methods. Only source-target alignments with results from both methods (SPD-HAO, TGMA-HAO and FBbt-HAO) are shown. Correspondences found by LOOM alone are further characterized as valid (overlooked during the manual alignment), mismatched (invalid correspondences) or other (errors; see text).

HAO. The small number of correspondences between all ontologies (totaling 11) suggests that many higher level arthropod classes are missing from one or more of the ontologies (e.g. male reproductive system, cuticle, nervous system) Although the lack of certain classes is likely a result of the varied purposes of each ontology, and thus the selection of classes to be included (below), identification of these missing classes will be important if extending the current ontologies is a common goal to increase their intercommunication. As reflected in this study, the use of domain experts in the process of identifying these classes will probably be critical.

Non-matches were generally due to the types of classes represented in the ontologies (differences specific to the taxon's anatomy, or based on developers' priorities/expertise) or difficulties in evaluating similarity between arachnid and insect classes. For example, specific wing veins, although uniform and easily defined in the single species *D. melanogaster* (at least in the wild-type; FBbt) and the family Culicidae (TGMA), were not able to be ontologically characterized for Hymenoptera (HAO) due to the immense diversity of wing venations represented by its members. Direct, 1:1 correspondences were typical, but on several occasions multiple classes from one ontology were found to be represented by only one class in HAO. Several different muscles characterized as separate classes in FBbt were found to represent one large, undivided muscle in HAO. Thus, all FBbt muscle classes identified as such were represented as a correspondence with one muscle class in HAO (e.g. HAO:0000332—first mesopleuro-mesonotal muscle—was found to be composed of the FBbt classes coxal tergal remotor muscle 48a, coxal tergal remotor muscle 48b, tergosternal muscle 47b, tergosternal muscle 47a and tergosternal muscle 47c). Furthermore, although all ontologies included at least some classes from CARO, the usage of this upper level ontology has been described as 'not . . . very consistent' (21) and was observed by us as well (Table 2). Whether CARO 2.0 (21) will ameliorate these issues is to be seen.

The following sections briefly describe the general types of classes that were either matched or were not found to

Database, Vol. 2013, Article ID bas057, doi:10.1093/database/bas057

Table 3. List of relations used in each ontology[a]

Hymenoptera (HAO)	Spiders (SPD)	Ticks (TADS)	Mosquitoes (TGMA)	*Drosophila melanogaster* (FBbt)
part_of	***part_of***	***part_of***	***part_of***	***part_of***
attached_to			*has_part*	*axon_innervates*
integral_part_of			*develops_from*	*connected_to*
				dendrite_innervates
				develops_directly_from
				develops_from
				electrically_synapsed_to
				fasciculates_with
				has_function_in
				has_part
				has_quality
				has_soma_location
				innervated_by
				innervates
				overlaps
				partially_overlaps
				releases_neurotransmitter
				secretes_hormone
				synapsed_by
				synapsed_to

Bold relations denote those shared by all. [a]Does not include *is_a*, *disjoint_from* and others that are implicit in the OBO format (http://www.geneontology.org/GO.format.obo-1_4.shtml).

have correspondences with HAO (see Figure 1 for additional details).

SPD versus HAO

Although both the HAO and SPD (Figure 1, purple diamonds) had many classes regarding external anatomy, each also had classes that were either domain specific or not yet addressed in the other ontology. Those that were matched pertained mostly to leg segments, some aspects of the reproductive system and higher level CARO classes. SPD did not contain muscle classes that were heavily characterized in HAO. Conversely, SPD had many classes associated with silk types and silk production, types of eyes, male secondary sexual organ (palp) anatomy and setae/sensory structures, all of which either do not exist in Hymenoptera or were characterized at a coarser level in HAO.

TADS versus HAO

TADS (Figure 1, green squares) is largely based on general tick anatomy described in (14). Since that text deals with multiple organ systems and structures, TADS has a broad base of classes. Of these systems, most of the correspondences found between TADS and HAO pertained to external skeletal structures and various tissues, organs or muscles. TADS classes that were not generally matched with those in HAO included specific tracheal system components, certain tissues and organs, nerves or secretory glands. These unmatched classes were either specific to ticks or were not yet characterized in the hymenopteran ontology.

TGMA versus HAO

TGMA (Figure 1, red octagons) covers mainly the external anatomy of adult and larval mosquitoes. It does characterize some internal structure (e.g. some apodemes and portions of the internal genitalia), but does not include muscles or many components of organ systems. The main overlap between HAO and TGMA was adult skeletal structures and other external sclerites, as well as certain leg and wing structures. TGMA classes that were not mapped to HAO usually involved specific structures found in mosquito larvae or eggs that are either not present in hymenopterans or have not been included in HAO due to its focus on adult anatomy. Others involved specific setae, setal patches or spicules, classes important in mosquito taxonomy/identification, but not present in the Hymenoptera ontology.

FBbt versus HAO

FBbt (Figure 1, blue triangles) covers a wide range of classes focused on the internal and external anatomy of the model

organism *D. melanogaster*. Classes that were generally mapped between FBbt and HAO were those describing external sclerites, leg and wing structures, certain tissues and muscles. FBbt contained many classes defining specific neurons and nervous system components, precursor cells and other cell types. FBbt also contains many classes for egg, embryonic and larval structures which are generally not found in HAO, the latter focusing mainly on adult anatomy.

Issues encountered while matching ontologies

During the matching process, several issues became apparent that hindered or may hinder both manual and algorithm-based methods. Some, like different levels of granularity, defined by how subdivided an ontology is [for instance, HAO has both 'anterior notal wing process' (HAO:0000120) and 'posterior notal wing process' (HAO:0000758), but not the general superclass 'notal wing process' that is found in FBbt (FBbt:00004584)], were often encountered and appear to represent a common difference among ontologies (thus the issue is not discussed here). Others, such as lacking text definitions, were mostly restricted to one or two of the ontologies.

Solely using algorithms that match through logically asserted relations would have been hampered by the structural heterogeneity (i.e. differences in where relations are applied) of the ontologies. As a basic proxy for structural similarity, we calculated the number of direct superclasses (also called direct parents or ancestors) that matched, related to the correspondences found (Table 2). Specifically, we calculated whether the direct superclasses on each side of the paired correspondences were also matched in the alignments—superclass matches were considered evidence of similar structure (i.e. relations made in similar ways/directions). Although crude, and possibly affected by differences in granularity (missing intermediate classes, etc.), we feel this to be a simple way to discretely view structural differences between the ontologies. The results showed that the local structure of these ontologies were quite different from each other: just looking at subsumption (*is_a* relations), TADS and HAO appear to have the most similar structure in relation to their correspondences (49 of the 63 *is_a* relations matched the same superclasses), while FBbt and HAO were the most dissimilar (only 93 of the 395 *is_a* relations matched the same superclasses) (Table 2). However, none has exactly the same structure, even though the classes in question are putatively congruent across the ontologies, leading us to consider that automated, relation-based alignments would have had difficulty identifying correspondences that were found using labels and definitions by us, the domain experts.

Several classes among the ontologies had misspelled labels in the 'Name' field. These included, for example, 'protharocic notal plate' (TGMA), 'adult accessroy nerve ROC' (TADS) and 'adult Gene's organ horm' (TADS) (corrected labels having 'prothoracic', 'accessory' and 'horn', respectively); even 'spermathecum' (FBbt), although ostensibly spelled correctly, is misapplied as being neuter singular rather than the correct feminine singular, 'spermatheca' (pl. 'spermathecae'). Because these and others are misrepresented, it may be difficult for some algorithms and non-domain experts to identify correspondences involving these labels, difficulties that are further compounded when misspelled classes lack good definitions.

Regarding definitions: while computer reasoning across ontologies is often accomplished through logically asserted relations, and not through text definitions, humans performing manual alignments or evaluating automated results often require some idea of a class's meaning. This is accomplished by understanding its definition. Unfortunately, true text definitions were not always fully represented (e.g. FBbt only has 43% of its classes represented by these definitions; Table 1), nor are they always represented in a useful way. From our own experience developing the HAO, attempts were made to have complete *genus-differentia* definitions for all classes that we created, and at least some definition for all classes regardless of format (e.g. definitions taken directly from cross-referenced classes in other ontologies were often adopted verbatim). Instances where a lack of good definitions hindered manual alignment were common, especially when dealing with the FBbt which, as stated above, lacks many definitions.

Another issue that may not be present in ontologies of more closely related organisms (e.g. mouse and human ontologies), but which arises when classes from disparate groups are being matched, is homonomy, or the use of the same name/label for different classes in different ontologies. Several instances of these were encountered during the matching process, such as 'radix' (HAO versus SPD), 'serrula' (HAO versus SPD), 'pedicel' (HAO versus SPD), 'alveolus' (HAO versus SPD), 'metatarsus' (HAO versus FBbt), 'lamina' (HAO versus FBbt) and 'flange' (HAO versus FBbt). For example, 'radix' in Hymenoptera refers to an area on the egg-laying device (ovipositor), while in spiders it refers to a structure in the male secondary sex organ (palp). Both are derived from the same descriptive word origin, but do not represent the same class concept. Although they did not significantly hinder the current analysis, the presence of homonyms could easily cause issues for automated matching algorithms (especially those based largely on lexical matches) and manual methods performed quickly without knowledge of the underlying differences between classes with the same labels (i.e. matching performed by non-domain experts).

Finally, a difficulty with developing ontologies for arthropods, especially insects that undergo complete metamorphosis (Holometabola or ~75% of all known life), is that they must take into account the anatomy of different

Database, Vol. 2013, Article ID bas057, doi:10.1093/database/bas057

life stages, i.e. the progression of morphological diversity throughout the organism's development. Many anatomical features are specific to only the egg, larva, pupa or adult, while others are applicable to two or more of the life stages. Often information about both larvae and adults (stages that can differ immensely in general morphology) is important and must be properly characterized, ontologically. However, this raises some considerations: do we divide the ontologies into one for adults and another for larvae or do we attempt to unify them by creating specific classes for each life stage? Each strategy has implications, but all the ontologies discussed here, when necessary, provide stage-specific classes interspersed with classes common to all of the life stages in one ontology. This approach is logical for maintaining all classes for an organism in one location, but it is not without issues. For instance, the HAO, although representing a holometabolous group, is mainly concerned with adult morphology and is unlikely to include many larva-specific classes in the near future; thus all classes, unless stated otherwise, are considered adult specific. The problems with this approach are (i) larval structures exist and may need to be incorporated later and (ii) each class has the potential duality of representing a general class and an adult class. Both these factors contributed to difficulties during alignment, because a dual class such as 'thoracic segment' in HAO could be aligned with either 'thoracic segment' in FBbt (the general class) or 'adult thoracic segment' in FBbt (the stage-specific class). Ultimately, we aligned these classes on a case-by-case basis depending on the number of general or stage-specific classes present. Another factor is that the prevalence of stage-specific classes will surely depend on the taxon that is being covered. While some taxa require extra classes for different life stages, others, like many arachnid and insect groups, change very little between life stages; they usually only develop reproductive organs or wings, but remain almost identical otherwise. In these cases, making every class have a stage-specific component (i.e. a juvenile and adult class for each structure) will certainly result in much more effort than is necessary to have a functioning ontology. In the future we will endeavor to create general classes first, then applying stage-specific classes as children/subclasses, as necessary. FBbt employs this approach (e.g. both 'larval thorax' and 'adult thorax' *is_a* 'thorax'), while the TGMA contains examples where stage-specific classes are related to a higher level (e.g. 'larval thorax' and 'adult thorax' *is_a* 'organism subdivision').

Comparison with LOOM

LOOM is a strictly lexical matching tool that compares the preferred names (labels) and synonyms of classes in each ontology to achieve an alignment (after standardized transformations of the text string). Superficially, it appeared that the algorithm was more successful at finding correspondences. This was true in a few cases where matches were found that were not identified during the manual process, likely occurring because of errors handling large amounts of data by domain experts and resulting in several valid correspondences being overlooked. However, upon further investigation, many of the correspondences found using LOOM alone were either found to be invalid matches (up to 65% of total correspondences in TGMA) or other errors. Furthermore, LOOM failed to recognize a number of exact lexical matches between the ontologies; the reason for the software overlooking these valid matches is unknown to us. Overall, the algorithm slightly improved some results, but many of its propositions were identified as invalid when evaluated by domain experts.

In contrast, many newly proposed correspondences were made directly and solely by us, the domain experts. These correspondences were not found using simple lexical matching methods and would not have likely been found by more sophisticated logic-based, reasoning methods because of structural differences between the ontologies. For example, many of the muscles aligned between the ontologies (especially between HAO and FBbt) had no ontological evidence for correspondence and were only discovered by looking at primary literature that had characterized the musculature of these organisms. Expert-based domain knowledge and reference to literature also aided in elucidating other types of classes. The use of human input, therefore, appears to be crucial for recognizing correspondences for difficult class concepts (especially those without similar labels) and vetting those found using algorithms.

Despite the marked increase in identifying correspondences using the manual method, we propose using both approaches since neither is perfect at finding all correspondences. Our results and those described in [8] suggest using both approaches together, allowing each to validate the other through a combination of lexical, structural and domain expert analysis.

Conclusions and future directions

Although these ontologies are not static and have evolved from their state presented here, the alignments described by us are important sets of correspondences and represent a baseline from which to work. We recognize that the content, structure and functionality of an ontology are related to (and derived from) the uses intended by those developing it, and the needs of the domain of interest. The preservation of this functionality is a major factor for its content and future utility. However, the potential need and benefit for communication between ontologies means that they cannot be developed solely in isolation. Thus, these correspondences should prove useful for extending and

harmonizing the ontologies and for guiding the formation of future ones for other groups of arthropods.

The results of this study are presently being considered to guide a common arthropod/insect anatomy ontology, spearheaded by the Phenotype Research Coordination Network group (http://phenotypercn.org). This base ontology should aid developers and domain experts who would like to adopt a common set of classes and their logical relations for this group of organisms, all of which have been evaluated and reconciled across the diversity of Arthropoda and Insecta. This would most likely require creating unified classes for each correspondence, all with computable definitions and reference to the ontologies involved. It may also be beneficial to create multiple base ontologies for different taxonomic levels, i.e. one for arthropods, insects and holometabolous insects (ones with complete metamorphosis, requiring stage-specific classes as discussed above), to relieve the need to create many unnecessary classes.

Acknowledgements

The authors thank Lars Vilhelmsen and Gary Gibson for invaluable advice on some aspects of Hymenoptera anatomy. They also thank Chris Mungall and two anonymous reviewers for comments and suggestions on an earlier version of this article.

Funding

The U.S. National Science Foundation (NSF) [grant number DBI-0850223]; the National Evolutionary Synthesis Center (NESCent) [NSF grant number EF-0905606]. Ideas explored herein also benefited from discussions fostered through the Phenotype RCN [NSF grant number EB-0956049].

Conflict of interest. None declared.

References

1. Smith,B., Ashburner,M., Rosse,C. *et al*. (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
2. Ramírez,M.J., Coddington,J.K., Maddison,W.P. *et al*. (2007) Linking of digital images to phylogenetic data matrices using a morphological ontology. *Syst. Biol.*, **56**, 283–294.
3. Topalis,P., Tzavlaki,C., Vestak,K. *et al*. (2008) Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase. *Insect Mol. Biol.*, **17**, 87–89.
4. Drysdale,R. (2001) Phenotypic data in FlyBase. *Brief. Bioinform.*, **2**, 68–80.
5. Yoder,M.J., Mikó,I., Seltmann,K.C. *et al*. (2010) A gross anatomy ontology for Hymenoptera. *PLoS One*, **5**, e15991.
6. Euzenat,J. and Shvaiko,P. (2007) *Ontology Matching*. Springer, Heidelberg.
7. Shvaiko,P. and Euzenat,J. (2008) Ten challenges for ontology matching. In: *Proceedings of the 7th Conference on Ontologies, Databases, and Applications of Semantics (ODBASE),* 11–13 November 2008. Monterey, Mexico, pp. 1164–1182.
8. Bodenreider,O., Hayamizu,T.F., Ringwald,M. *et al*. (2005) Of mice and men: aligning mouse and human anatomies. In: *Proceedings of the American Medical Informatics Association (AIMA) Annual Symposium,* Washington DC, USA, pp. 61–65.
9. Hayamizu,T.F., de Coronado,S., Fragoso,G. *et al*. (2012) The mouse–human anatomy ontology mapping project. *Database*, **2012**, article ID bar066; doi:10.1093/database/bas066.
10. Haendel,M., Gkoutos,G.V., Lewis,S.E. *et al*. (2009) Uberon: towards a comprehensive multispecies anatomy ontology. *Nat. Precedings*, doi:10.1038/npre.2009.3592.1.
11. Mungall,C.J., Torniai,C., Gkoutos,G.V. *et al*. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
12. Stevens,R., Goble,C.A. and Bechhofer,S. (2000) Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.*, **1**, 398–414.
13. Seltmann,K.C., Yoder,M.J., Mikó,I. *et al*. (2012) A hymenopterists' guide to the Hymenoptera Anatomy Ontology: utility, clarification, and future directions. *J. Hymenopt. Res.*, **27**, 67–88.
14. Ubick,D.P., Paquin,P., Cushing,P.E. *et al*. (2005) *Spiders of North America: An Identification Manual*. American Arachnological Society.
15. Sonenshine,D.E. (1991) *Biology of Ticks*, Vol. 1. Oxford University Press, Oxford, New York.
16. Demerec,M. (1994) *Biology of Drosophila*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
17. Zalokar,M. (1947) Anatomie du thorax de *Drosophila melanogaster*. *Rev. Suisse Zool.*, **54**, 17–53.
18. Harbach,R.E. (2011) Mosquito Taxonomic Inventory. http://mosquito-taxonomic-inventory.info/ (19 May 2011, date last accessed).
19. Ghazvinian,A., Noy,N.F. and Musen,M.A. (2009) Creating mappings for ontologies in biomedicine: simple methods work. In: *AMIA. Annual Symposium (AMIA 2009),* San Francisco, CA.
20. Haendel,M., Neuhaus,F., Osumi-Sutherland,D. *et al*. (2008) CARO—the common anatomy reference ontology. In: Burger,A., Davidson,D. and Baldock,R. (eds), *Anatomy Ontologies for Bioinformatics: Principles and Practice*. Springer, Heidelberg.
21. Osumi-Sutherland,D. (2011) CARO 2.0. In: *Proceedings of the International Conference on Biomedical Ontology,* Buffalo NY, USA.