Lessons Learned in Aligning Multiple Anatomical Ontologies across Different Species

Songmao Zhang[1], Ph.D., Olivier Bodenreider[2], M.D., Ph.D.

**[1]Institute of Mathematics, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing, P. R. China
[2]U.S. National Library of Medicine, NIH, Bethesda, Maryland, USA**
[1]smzhang@math.ac.cn, [2]olivier@nlm.nih.gov

Objective: To examine the entent to which five anatomical ontologies for different species, human, mouse, fly, worm and fish, can be aligned automatically. Methods: Pairwise mappings among the five ontologies were created using a combination of lexical alignment and structural validation techniques. A manual review of a limited number of mappings was performed to identify the limitations of this approach. Results: Less than 5% of the concepts from a given ontology were identified in another ontology for 15 of the 20 pairs of ontologies investigated. Conclusions: The precision and recall are generally low. The structural similarity required for the validation of lexical mappings would benefit from being tightened. Semantic validation would have prevented a small number of mismatches. Multiple mappings require disambiguation by a domain expert. Idiosyncrasies in naming and in knowledge representation account for a significant proportion of missed matches. Overall, aligning anatomical ontologies across species remains difficult.

INTRODUCTION

Biomedical research relies on model organisms [1]. The functional description of experimental data has benefited from the standardization supported by initiatives such as the development of the Gene Ontology (GO) [2], a controlled terminology for the functional annotation of gene products across species embraced by most model organism communities. Similar efforts have been initiated for anatomy. The Foundational Model of Anatomy (FMA) has been proposed as an anatomical ontology for vertebrates [3], in addition to humans, but has not been widely adopted by biologists yet. The Common Anatomy Reference Ontology (CARO) is "being developed to facilitate interoperability between existing anatomy ontologies for different species" [4]. The current version (1.3) is comprised of concepts for 46 high-level anatomical entities. While promising, CARO is currently not usable in practice for the detailed annotation of anatomical structures across model organisms.

Aligning ontologies, especially anatomical ontologies, is a nontrivial task [5]. It has been proposed for several years as one task of the Ontology Alignment Evaluation Initiative (OAEI) challenge [6], under the rubric "expressive ontologies". Most alignment systems tested did not perform well, when at all, in the early years of the competition. One problem common to most alignment systems was the use of generic alignment systems for aligning specialized terminologies, yielding poor recall and precision [7].

Aligning anatomical ontologies not only across mammalian species, but also across species exhibiting fundamental differences from an evolutionary perspective is, of course, more challenging. In addition to genuine anatomical differences among species, different communities might have selected different names to denote similar anatomical entities. Despite these differences, establishing correspondences among anatomical concepts across ontologies for multiple species is important for comparative genomics and, more generally, for translational research, where data integration plays a fundamental role [8].

The objective of this study is to examine the extent to which five anatomical ontologies for species including human, fly, mouse, worm and fish can be aligned using automatic techniques. This paper reports on the lessons learned in performing this alignment and discusses false positives, false negatives, ambiguous mappings and differences among anatomical ontologies.

MATERIALS

Anatomical ontologies are comprised of concepts representing anatomical entities and their interrelations. Two major relationships form the backbone of anatomical ontologies. The *is-a* relationship is used to represent anatomical taxonomies and the *part-of* relationship defines mereologic (part-whole) relations among anatomical entities. The five anatomical ontologies under investigation in this study are indeed organized around *is-a* and *part-of* relations. These ontologies were downloaded from the Open

Biomedical Ontologies web site [9] on October 25, 2007.

The **Foundational Model of Anatomy** (FMA) is an evolving ontology developed by University of Washington, whose objective is to conceptualize the physical objects and spaces that constitute the human body [3]. Characteristics: 75,147 concepts; one preferred term per concept; 45,118 synonyms; *is-a* and *part-of* relations.

Adult Mouse Anatomical Dictionary (MA) is a structured controlled vocabulary describing the anatomical structure of the adult mouse [10], developed at the Jackson Laboratory as part of the Mouse Genome Database. Characteristics: 2,745 concepts; one preferred term per concept; 329 synonyms; *is-a* and *part-of* relations. MA is listed under the name *Mouse adult gross anatomy* in OBO.

Fly Anatomy (FA) is the anatomical and developmental vocabulary developed in conjunction with FlyBase, which is a collection "of genetic and genomic data on the model organism *Drosophila melanogaster* and the entire insect family Drosophilidae" [11], developed by the FlyBase Consortium. Characteristics: 6,024 concepts; one preferred term per concept; 1,235 unique synonyms (and 1,483 shared synonyms, not used in our study as they introduce ambiguity); *is-a* and *part-of* relations. FA is found under the name *Drosophila gross anatomy* in OBO.

Worm Anatomy (WA) is the anatomical and developmental vocabulary developed in conjunction with WormBase, which is a "model organism database for *Caenorhabditis elegans* and other related nematodes" [12], developed by the WormBase Consortium. Characteristics: 6,301 concepts; one preferred term per concept; 1,391 unique synonyms; *is-a* and *part-of* relations. WA exhibits some unusual features. Many concepts names in WA are shared by several concepts (e.g., 79 concepts share the same name *mu_bod*) and a very large number of the concepts (2,616) do not have any *is-a* or *part-of* connections to any other concepts. Such "dangling" concepts were not used in our study as hierarchical relations are a key element of our alignment technique. WA is found under the name *C. elegans gross anatomy* in OBO.

Zebrafish anatomical ontology (ZFA), developed by a consortium of researchers, is part of the Zebrafish Model Organism Database, which is "a web based community resource and a model organism database that implement the curation of zebrafish genetic, genomic and developmental data" [13]. Characteristics: 2,132 concepts; one preferred term per concept; 1,064 unique synonyms; *is-a* and *part-of* relations; 49 "dangling" concepts (ignored). ZFA is listed under the name *Zebrafish anatomy and development* in OBO.

METHODS

In order to align the five anatomical ontologies, we create ten direct, pairwise alignments. Each pairwise alignment is obtained through a combination of lexical and structural techniques. More precisely, we first compare terms across ontologies lexically in order to identify one-to-one concept matches. The second step is the validation of lexical matches using structural information. The interested reader is referred to [14] for details about our method.

Aligning ontologies pairwise

The **lexical alignment** compares two ontologies at the term level, by exact match and after normalization. Both preferred terms and synonyms in the two ontologies are used in the alignment. For example, the concepts *heart valve* in MA and *Cardiac valve* (synonym: *Heart valve*) in FMA are identified as a match. Moreover, synonymy information from external domain resources is used to identify additional matches. For example, *tooth pulp* in ZFA and *dental pulp* in MA, although lexically different, are considered a match because they name the same anatomical concept in the Unified Medical Language System® (UMLS®) [15].

The **structural validation** first acquires the inter-concept hierarchical relations, *is-a* and *part-of*, and their inverses, *inverse-isa* and *has-part*, respectively. Missing relations are generated through complementation, augmentation and inference techniques [14]. Once all relations are represented consistently, the structural alignment is applied to the matches resulting from the lexical alignment in order to identify similar hierarchical paths to other matches across ontologies. For example, the matching concepts *heart valve* in MA and *Cardiac valve* in FMA exhibit similar hierarchical paths to other matches in these two ontologies, including paths to *Heart* (*part-of*) and to *Aortic valve* and *Mitral valve* (*inverse-isa*). Such structural similarity is used as **positive evidence** for the alignment. Instead of similar paths, one match may exhibit paths to other matches in opposite directions in the two ontologies. Such paths suggest a structural conflict across ontologies. For example, in MA *pericardial cavity* stands in a *has-part* relation to *pericardium*, while in the FMA *Pericardial cavity* is defined as *part-of Pericardial sac*, which is *part-of Pericardium*. These conflicts are used as **negative evidence** for the alignment, indicating the semantic incompatibility between concepts across ontologies in spite of their lexical resemblance.

Aligning the five anatomical ontologies

We applied the alignment techniques presented above to all pairwise combinations of the five anatomical ontologies, resulting in ten pairwise alignments. Lexical mappings supported by at least one piece of positive structural evidence were considered a match. We excluded mappings not supported by any structural evidence, as well as those exhibiting negative evidence.

RESULTS

The results of the ten pairwise alignments among the five anatomical ontologies are summarized in **Table 1**, along with details about lexical alignment and structural validation. For example, in the mapping between MA and ZFA, 238 matches were identified through lexical alignment, including 37 identified through synonymy in the UMLS. These 238 matches represent 8.67% of the 2,745 concepts in MA and 11.16% of the 2,132 concepts in ZFA. Of these 238 matches, 212 (89.08%) were supported by positive structural evidence and finally reported as matches. The 26 matches not supported by structural evidence were ignored. No conflicts (negative structural evidence) were identified in this alignment.

DISCUSSION

The main issues we encountered in mapping among five anatomical ontologies across different species can be grouped into the following categories: false positives, false negatives, multiple ambiguous mappings, and differences among ontologies. This section provides a limited error analysis and outlines solutions for addressing these issues.

False positives

The presence of false positive mappings can be traced back to limitations in the mapping techniques used, including lack of semantic validation, insufficient structural validation and the influence of acronyms on lexical alignment.

We used **semantic validation** in previous alignment studies, primarily in order to distinguish between anatomical and non-anatomical entities across ontologies. In the present study, because the five ontologies under investigation are all restricted to the anatomical domain, we thought that the semantic validation would not be necessary. In fact, our limited review of some of the mappings led us to conclude that semantic validation would still be appropriate in this case. For example, in the mapping {FA: *accessory mesothoracic neuromere* (synonym: *ovoid*), FMA: *Ovoid*}, *accessory mesothoracic neuromere* in

FA *is-a neuromere* which *is-a ganglion*, while *Ovoid* in FMA *is-a Volume* which *is-a Dimensional entity*. There are no dimensional concepts in FA, so by specifying that anything in FA is disjoint with the top-level concept *Dimensional entity* in FMA, the mismatch between the two concepts could be detected automatically.

Insufficient structural validation can be blamed for some false positive matches. In fact, as mentioned earlier, structural validation only requires one shared path between lexical matches across ontologies in order to validate the mapping. We noticed that, in some cases, the shared path used as positive evidence involves only high-level concepts such as the root of some hierarchy. Take the mapping {WA: *axis*, FMA: *Axis*} for example. In WA, *axis* is defined as "spatial axis" and stands in an *is-a* relation to *anatomy*. Its three children are: *anterior-posterior*, *dorsal-ventral*, and *left-right*. On the other hand, in the FMA, *Axis is-a Cervical vertebra*, which *is-a Vertebra*, which *is-a Irregular bone*, which *is-a Organ*, which *is-a Anatomical structure*. The mapping received positive evidence in the structural validation phase, because in both systems *axis* is ultimately classified under *Anatomical structure* (mapped to *anatomy* in WA). While the mapping of spatial axis to a cervical vertebra is an obvious mismatch, this false positive was not detected by structural validation. In fact, the requirement that at least one hierarchical path be shared across ontologies for this mapping is insufficient, as it should exclude trivial shared hierarchical paths such as direct relations to the root (or to high-level concepts) of the anatomy hierarchy.

The presence of **acronyms in concept names** is also responsible for some false positive mappings. For example, {ZFA: *posterior macula* (synonym: *pm*), FA: *proximal medullary amacrine neuron* (synonym: *Pm*)} were identified as a match through synonymy. Both concepts are *part-of nervous system* and *whole organism*, so the mapping is supported by positive evidence. However, *posterior macula* in ZFA is defined as "patches of thickened, pseudostratified epithelium of the inner ear…" and stands in a *part-of* relationship to *inner ear*. In FA, *proximal medullary amacrine neuron is-a neuron* and stands in a *part-of* relationship to *brain* and *nerve*.

False negatives

As noted in previous studies [14], in some cases, differing knowledge representation strategies are responsible for the lack of shared hierarchical paths for equivalent concepts across ontologies. For example, {WA: *body wall musculature*, FMA: *Musculature of body wall* (synonym: *Body wall musculature*)} do not share any paths, but are nonetheless equivalent concepts. In WA, *body wall musculature is-a muscu-*

lar system, which *is-a Organ system*, and it has child *head muscle* and has *striated muscle* as part. In FMA, *Musculature of body wall* is *part-of Body wall* and *is-a Set of muscles of subdivision of trunk*. The two concepts do not share any hierarchical links to other concepts across systems due to differences in knowledge representation. Devising automatic methods for assessing such mappings automatically is extremely challenging.

Multiple ambiguous mappings

The alignment technique we developed is expected to identify point-to-point equivalences across ontologies. Multiple mappings occur when more than one concept from one ontology is mapped to one concept from another ontology. Such ambiguous mappings are indicative of an error, because, within one ontology, concepts generally denote distinct entities. A significant number of multiple mappings were identified in this experiment. For example, among 1,568 MA-FMA mappings, 138 MA concepts (8.80%) were mapped to more than one FMA concept and 21 FMA concepts (1.34%) mapped by more than one MA concept. Overall, the proportion of ambiguous mappings ranges from 0 to almost 10%. Mappings to the FMA tend to have higher proportions of ambiguous mappings.

Disambiguation of multiple mappings is required in order to select one valid mapping among the several mappings identified for one concept. However, disambiguation is often difficult due to a lack of detailed knowledge represented in the ontologies. For example, both *pharynx* and *esophagus* in FA were mapped to *pharynx* in WA, which has *esophagus* as a synonym. Both mappings received positive structural evidence because all three concepts are *part-of organ system* and *alimentary system*. In FA, both *esophagus* and *pharynx* are described as *part-of foregut*. *Esophagus* has one child *embryonic esophagus*, while *pharynx* is a leaf concept. In WA, *pharynx is-a organ*, *part-of digestive tract* and has no children. In the absence of distinctive relations to other concepts in the two ontologies, domain expertise is required to clarify the differences between esophagus and pharynx in fly and worm.

Differences among ontologies

One of the differences explaining the limited number of equivalent concepts identified across ontologies is the **difference in granularity**. As shown in other studies [16], some 60% of the anatomical concepts in the FMA differ from their parent concept(s) solely by laterality, i.e., most often by the presence of "left" or "right" in the concepts name. Since some ontologies purposely avoid representing laterality information for paired anatomical structures, failure to identify mappings for anatomical structures represented at this level of granularity does not constitute a limitation of the alignment system. In fact, among the ontologies under investigation in this study, FMA, MA and WA represent laterality information, while FA and ZFA do not. For example, in ZFA, *nasal artery* is not further classified. In MA there are two more specific nasal arteries, *dorsal nasal artery* and *ventral dorsal nasal artery*. In FMA, the classification is finer-grained, further classifying *Dorsal nasal artery* into *Left dorsal nasal artery* and *Right dorsal nasal artery*. Most fine-grained concepts were not identified automatically in our alignments.

The use of lexical alignment as the first step of the alignment process presupposes that concept names are amenable to natural language processing techniques, including edit distance and normalization. **Idiosyncrasies in naming** generally defeat the lexical alignment techniques. For example, concepts names specific to WA include *mu_bod* for "cell of the body wall muscles", as well as *lineage name\: MS.pppppp* and *Earaa*. In such cases, domain expertise is required to distinguish between anatomical entities for which a correspondence could be found in other ontologies under a different name, and concepts specific to a given species (e.g., specific cell lines).

Finally, there are **genuine anatomical differences among species**, including the presence of fins in fish (ZFA), wings in fly (FA) and whiskers in mouse (MA). Therefore, wing-related concepts from FA (e.g., *wing*, *ventral wing*, *wing margin*, *wing nerve*, *wing blade*) cannot be expected to be mapped to any concepts in any of the four other ontologies. Subtle structural differences can also be taken into account (e.g., the fact that mice have not one but five prostates [17]).

CONCLUSIONS

We studied the automatic alignment of anatomical ontologies for five different organisms along the evolutionary spectrum. Only a limited number of correspondences could be identified among ontologies. While 57% of MA concepts were identified in FMA, less than 5% of the concepts from a given ontology were identified in another ontology for 15 of the 20 pairs of ontologies investigated. With the exception of MA-FMA, the precision and recall are generally low.

The structural similarity required for the validation of lexical mappings would benefit from being tightened. Semantic validation based on disjoint top-level categories would have prevented a small number of mismatches. Multiple mappings require disambiguation by a domain expert. Idiosyncrasies in naming and in knowledge representation (including differences in

granularity) account for a significant proportion of missed matches.

Overall, aligning anatomical ontologies across species remains difficult. The evaluation of such studies suffers from the absence of gold standard alignments for most pairs of anatomical ontologies. Given the limited performance of automatic alignment techniques on these five anatomical ontologies, the automatic mapping would be best used to bootstrap a mapping curated by domain experts.

References

1. Beller M, Oliver B. One hundred years of high-throughput Drosophila research. Chromosome Res 2006;14(4):349-62
2. Blake JA, Bult CJ. Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform 2006;39(3):314-20
3. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478-500
4. Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee PM, Jr JLVM, Mungall CJ, et al. CARO — The Common Anatomy Reference Ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy Ontologies for Bioinformatics. Principles and Practice: Springer; 2008. p. 327-350
5. Euzenat J, Shvaiko P. Ontology matching. New York: Springer; 2007
6. Ontology Alignment Evaluation Initiative: http://oaei.ontologymatching.org/
7. Zhang S, Bodenreider O. Identifying mismatches in alignments of large anatomical ontologies. AMIA Annu Symp Proc 2007:851-855
8. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2008. Methods Inf Med 2008;47(Suppl 1):91-99
9. Open Biomedical Ontologies (OBO): http://www.obofoundry.org/
10. Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biology 2005;6(3):R29
11. Wilson RJ, Goodman JL, Strelets VB. FlyBase: integration and improvements to query tools. Nucleic Acids Res 2008;36(Database issue):D588-93
12. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, et al. WormBase: new content and better access. Nucleic Acids Res 2007;35(Database issue):D506-10
13. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, et al. The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Res 2006;34(Database issue):D581-5
14. Zhang S, Bodenreider O. Experience in aligning anatomical ontologies. International Journal on Semantic Web and Information Systems 2007;3(2):1-26
15. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32 Database issue:D267-70
16. Golbreich C, Zhang S, Bodenreider O. The foundational model of anatomy in OWL: Experience and perspectives. Web Semantics: Science, Services and Agents on the World Wide Web 2006;4(3):181-195
17. Travillian RS, Gennari JH, Shapiro LG. Of mice and men: Design a comparative anatomy information system. AMIA Annu Symp Proc 2005:734-748

Table 1. *Mappings obtained among five anatomical ontologies (ten pairwise alignments).*

Pair of ontologies	Lexical alignment				Structural validation					
	Number of mappings		Mappings / 1st ont.	Mappings / 2nd ont.	Positive evidence		No evidence		Negative evidence	
	Total	through UMLS								
MA-FMA	1,568	75	57.12%	2.09%	1,459	93.05%	105	6.69%	4	0.26%
ZFA-FMA	522	21	24.48%	0.69%	458	87.74%	62	11.88%	2	0.38%
FA-FMA	198	35	3.29%	2.63%	116	58.59%	80	40.40%	2	1.01%
WA-FMA	86	23	1.36%	0.11%	65	75.58%	19	22.09%	2	2.33%
MA-ZFA	238	37	8.67%	11.16%	212	89.08%	26	10.92%	0	0
MA-FA	90	18	3.28%	1.49%	66	73.34%	21	23.33%	3	3.33%
MA-WA	38	10	1.38%	0.60%	29	76.32%	9	23.68%	0	0
ZFA-FA	112	15	5.25%	1.86%	100	89.29%	12	10.71%	0	0
ZFA-WA	88	8	4.13%	1.40%	45	51.14%	43	48.86%	0	0
FA-WA	88	5	1.46%	1.40%	46	52.28%	40	45.45%	2	2.27%