Cross-species Mapping between Anatomical Ontologies: Terminological and Structural Support

Sarah Luger , Stuart Aitken and Bonnie Webber
The University of Edinburgh, School of Informatics
2 Buccleuch Place, Edinburgh EH8 9LW, UK
sluger@inf.ed.ac.uk, stuart@aiai.ed.ac.uk,bonnie@inf.ed.ac.uk

INTRODUCTION

Anatomical ontologies play an increasingly important role in indexing data, including gene expression, in model organisms. However, the terms used to name anatomical concepts (tissues) differ between communities, and the ontologies of model organisms differ from one another in ways that do not correspond to differences between the organisms themselves [Aitken, Webber and Bard, 2004]. We argue that the analysis of both terminology and structure is needed to support the alignment of anatomical ontologies across species, and propose automated methods for alignment.

The current work extends the terminological and ontological analysis of [Zhang, Mork and Bodenreider, 2004] through lexical analysis techniques that exploit the conceptual modelling adopted in the anatomies of model species. While Zhang et al. focus on aligning different anatomical ontologies for a single species (human), the overall aim of our project (www.xspan.org) is to establish alignments between anatomical ontologies for different species (C.elegans, drosophila, zebrafish and mouse). The baseline, to which linkages based on biological expertise and evidence from gene expression and cell-type data will be added, uses both structural and lexical clues. Model species anatomies differ from human anatomies in that some consist only of *part-of* relations (as opposed to both *part-of* and *is-a*), and none currently use complex logical definitions. As a consequence, the terms used to name concepts are central to both human and machine understanding of the anatomies.

METHODS

We employ a two-step approach, in which we first use lexical methods to normalize the terms in the ontologies in an effort to correctly align the mappings, and then test these language-based mapping judgments with an analysis of structural similarity. Lexical methods are used step-wise to normalize the terminology in the ontologies, in an effort to limit the effect of different descriptive styles. The methods include normalising spelling between English and American variants, reducing words to comparable forms using a stemmer and a lemmatizer, and removing stop words to ensure that only content words are compared. Each resulting descriptor is then treated as a set, to eliminate the effect of word order variation. There is one significant feature of the anatomical ontologies of certain model organisms that requires special handling: Only the *sequence of node labels* on the path from the root to node guarantees uniqueness, while in other ontologies, the *label* on a node constitutes its unique identifier. After lexical normalization, unique identifiers are compared two species at a time. Two methods of lexical alignment are examined, in the first, node label anchors with greater than 66% similarity are identified. These are then tested for structural consistency. A second lexical comparison is then run using the terms that make up a root to node path description of an anatomical part. Thus, this second alignment technique brings some structural information into consideration. Again, matches with a similarity score above a threshold are tested for structural consistency. Following Zhang, we seek *anchors* – these are concepts exhibiting lexical-level alignment between species. The structural connection between two anchors in the ontology is called a *bee-line*, and the types of links on a bee-line is usually significant. Due to the inconsistent treatment of *is-a* and *part-of* in the model species ontologies, structural similarity is evaluated by considering the ontologies as graphs with directed but unlabelled edges.

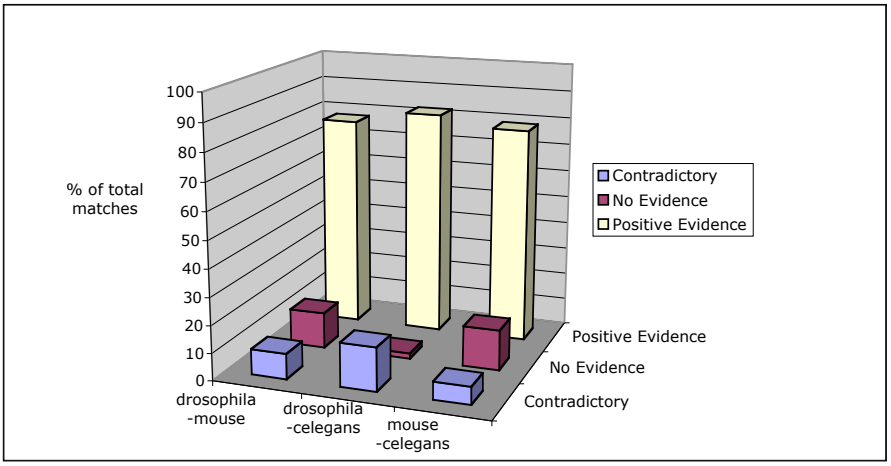

Figure 1: Structural evidence for lexical mappings

RESULTS

Using the anchor-based analysis of Zhang, we are able to assess the structural support for the proposed lexical matches. Results show that from 77-82% of lexical mappings have support from the ontology (indicated by the Positive Evidence bars in Figure 1) for the node-based comparison at 66% similarity. Less than 16% of proposed mappings have either no evidence for or against, or are contradictory[1] across the three comparisons drosophila-mouse, drosophila-celegans and mouse-celegans. For the path-based lexical mapping, which was tested at 75% and 85% similarity, the number of contradictory matches was reduced to 0, and the overall number of matches was also significantly reduced.

NEXT STEPS

After manual review, the lexically-based mappings between anatomies will be made publicly-accessible, in conjunction with expert and sequenced-based evidence, in a database of potential homologies and analogies between tissues. We shall seek to improve the precision of the lexical matches through the use of the synonym, abbreviation and lineage information that is contained in the model organism ontologies. We hypothesize that this will increase the recall and precision of the comparison. Since these synonyms only augment their original data set, another logical lexical step is using external synonyms from an anatomical reference source to increase the potential matches between species. While this may increase the number of false positive matches, the increase in data may allow insights into fine-tuning the optimal percentage of similarity between model species.

REFERENCES

Aitken, J.S., Webber, B.L. and Bard, J.B.L. Part-of Relations in Anatomy Ontologies: A Proposal for RDFS and OWL Formalisations. *Proc PSB 04*, 9:166-177(2004)

Aitken, J.S., Korf, R.,Webber, B.L. and Bard, J.B.L. COBrA: A Bio-ontology Editor. *Submitted to Bioinformatics*

Zhang, S. and Bodenreider, O. Investigating Implicit Knowledge in Ontologies with Application to the Anatomical Domain *Proc PSB 04* 9:250-261(2004)

Zhang, S., Mork, P. and Bodenreider, O. Lessons Learned from Aligning Two Representations of Anatomy. To be published in KR-MED (2004).

[1] A contradiction exists if a child node **B**, of a node **A** (where **A** and **B** are in ontology 1), maps to any node **B'** which is a parent of the node **A'** that **A** maps to (where **A'** and **B'** are in ontology 2).