

Abstract

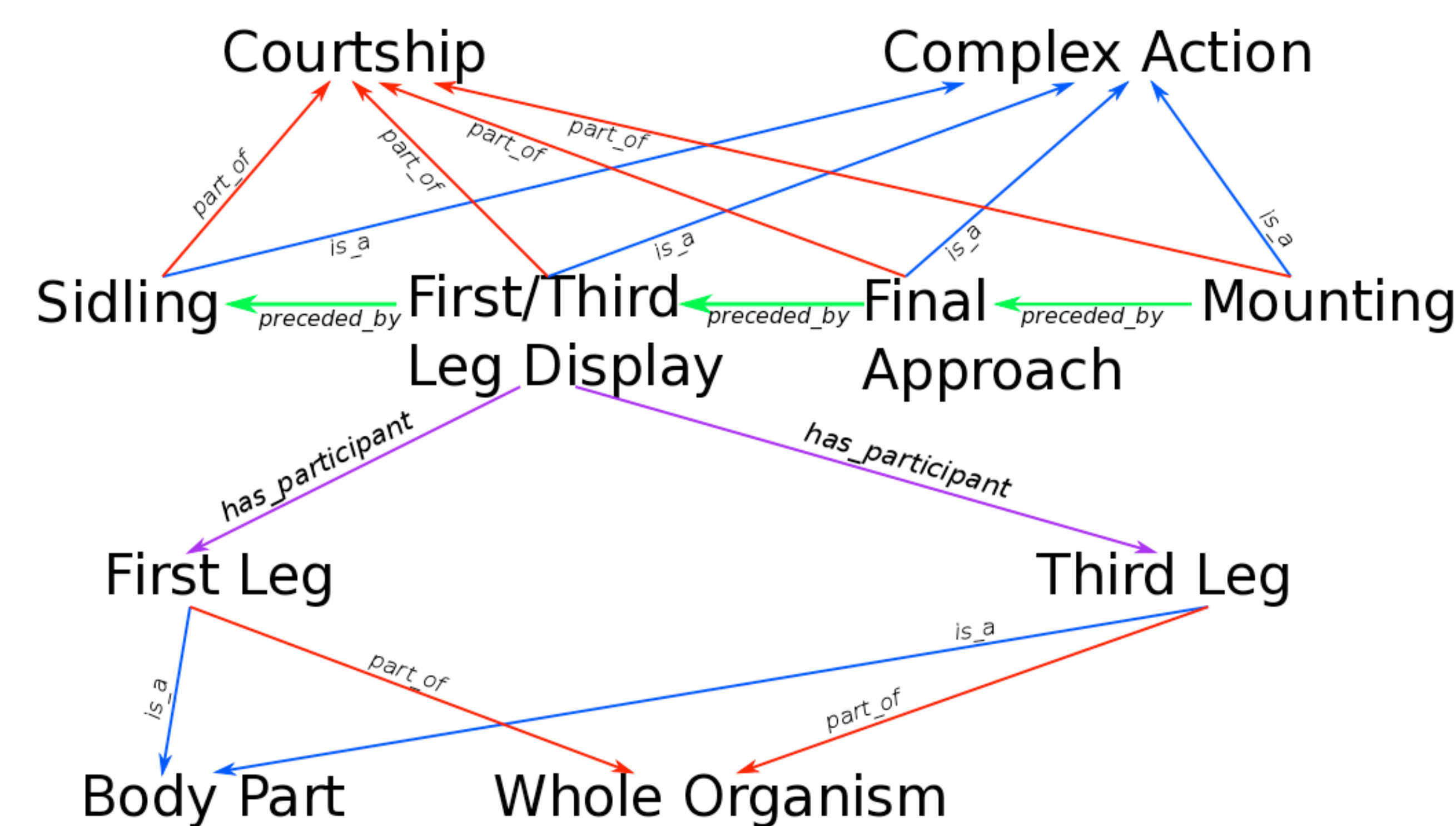
Phylontal is a simple approach to generating alignments of ontologies or constructs based on controlled vocabularies that reflect a phylogeny. The method seems to require only a downpass (tips to root) and thus suggests that constructing alternative alignments from multiple or uncertain phylogenies may be relatively inexpensive.

Introduction

Traditionally, phenotype data is captured using text descriptions. Even when coded in a character matrix, the states are generally defined either textually, or in some cases diagrammatically.

Ontologies have become an increasingly popular way to capture phenotype data in a form that machines can use to deduce conclusions from data using rules of inference (aka reasoning).

Ontologies consist of sets of terms that represent concepts of interest, generally either physical entities (e.g., leg, femur) or temporal processes (e.g., leg waving). These terms are linked by formally defined relations.



Fragment of *Habronattus* courtship ontology¹

As the number of ontologies has grown, so has the need to identify areas of overlap between two or more ontologies. One approach to identifying overlap is by finding terms and relations that are either identical between the ontologies, or for which correspondences can be identified. This approach is commonly called ontology matching² and the resulting sets of correspondences between parts of the matched ontologies is referred to as an alignment. A number of methods and tools have been developed over the past decade.

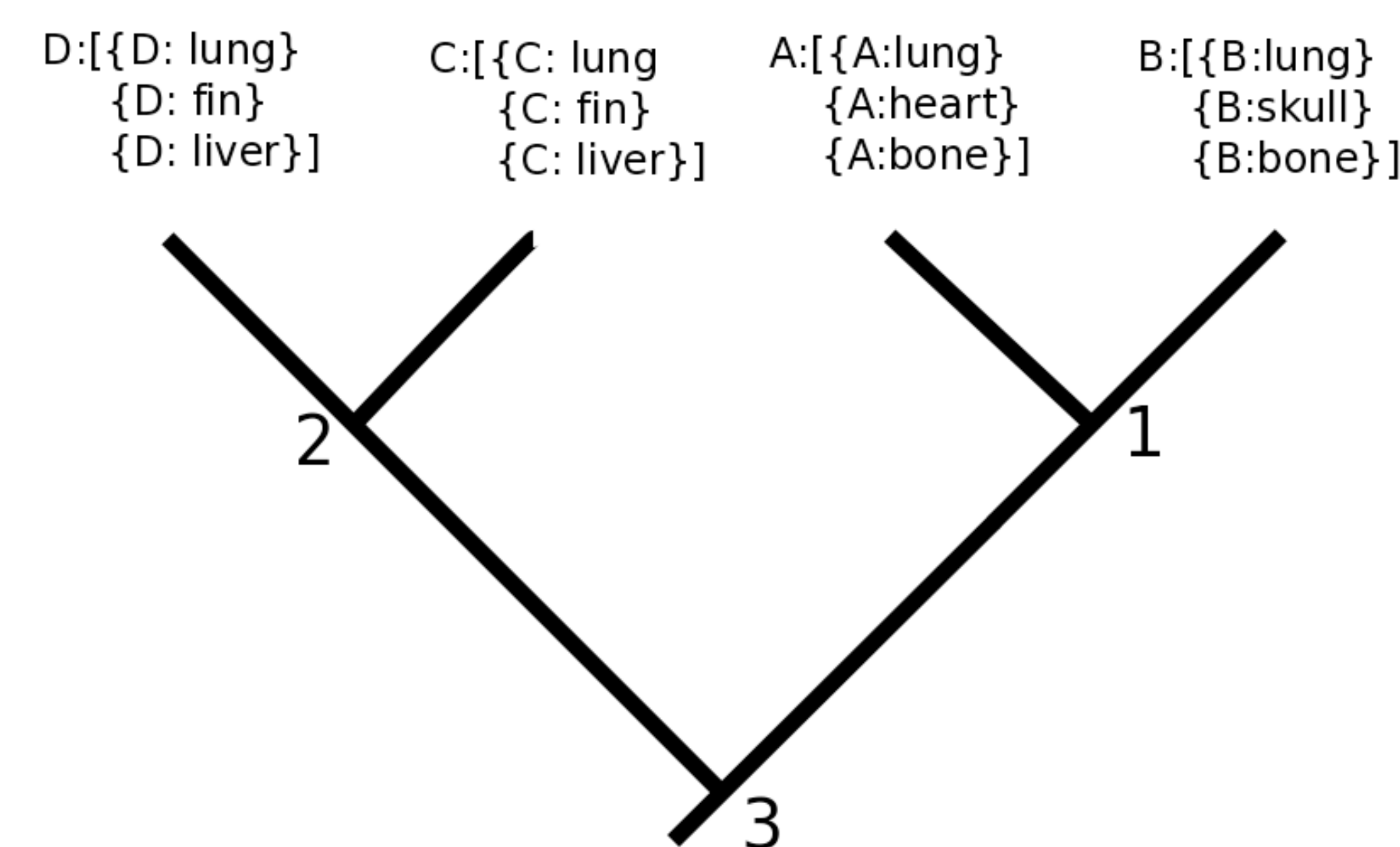
These methods typically use a combination of lexical matching and matching terms that have similar relations to terms already matched.

The Phylontal Method

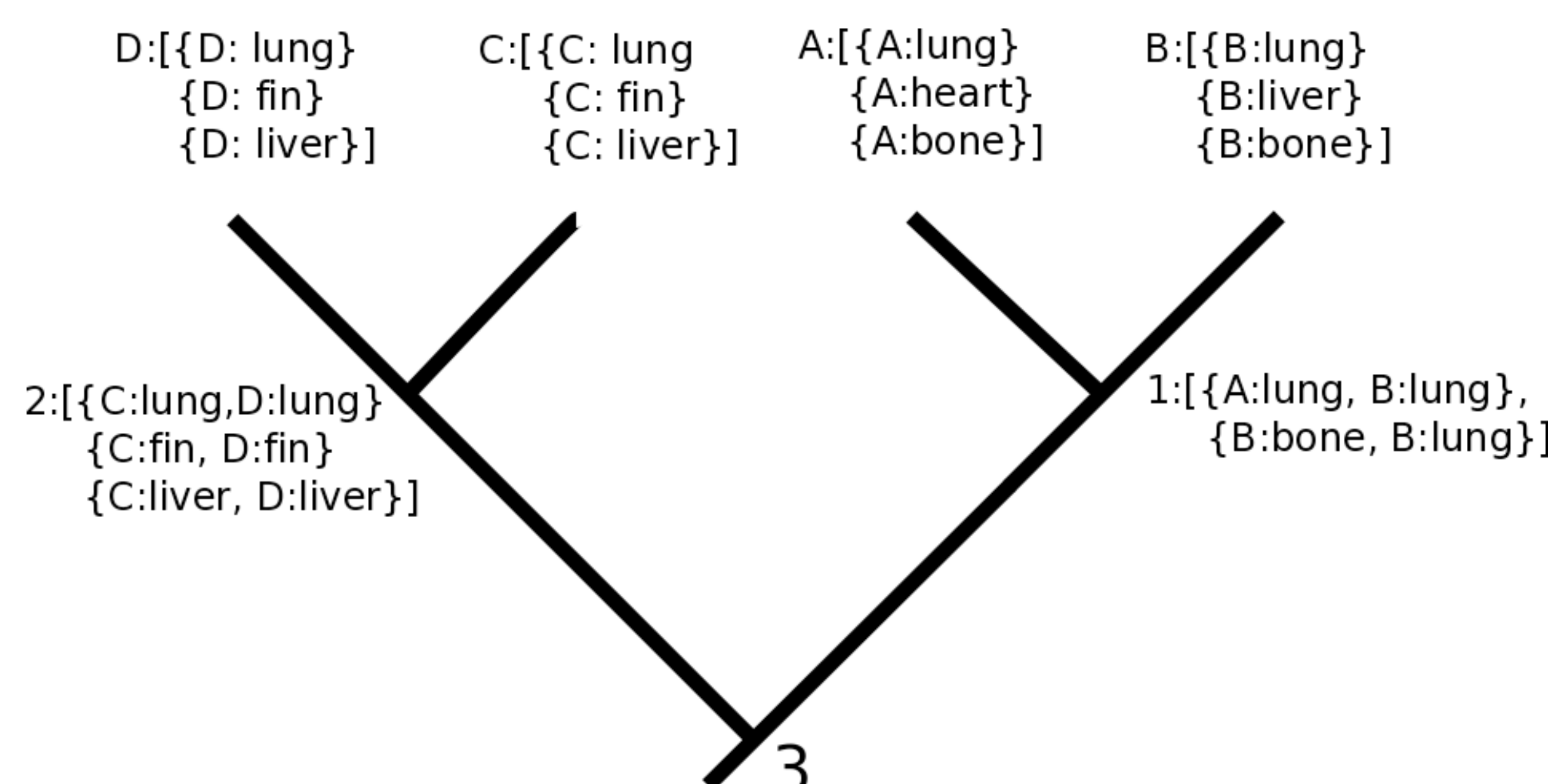
Phylontal is a method for identifying correspondences between terms in multiple ontologies and distinguishing homologous correspondences from those that are not homologous. The method, described in more detail in the middle panel, is based on pairwise alignment of ontologies in the course of a downward pass through a phylogeny.

Most non-evolutionary approaches create alignments that consist of pairs of corresponding terms between a single pair of ontologies. Another alignment tool for multiple taxa, Homolont³ creates sets of aligned terms, with each taxon contributing no more than one term to each set. Homolont's developers have also created a vocabulary of terms that describe correspondences, both homologous and non-homologous⁴.

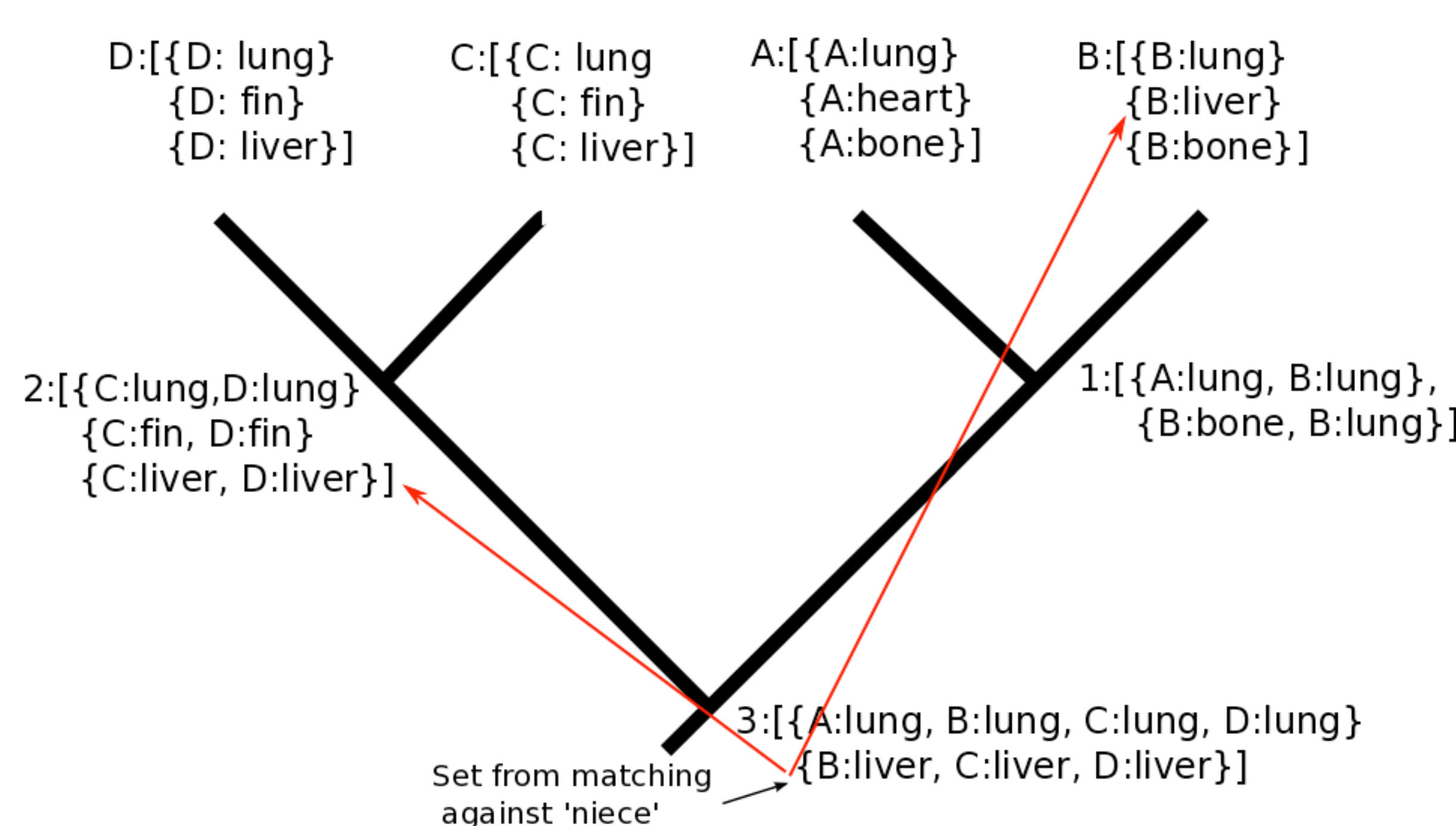
Phylontal creates collections of sets of terms that are associated with one (homologous correspondence) or more (non-homologous correspondence) nodes on the tree. The following is the procedure for constructing sets of homologous terms.



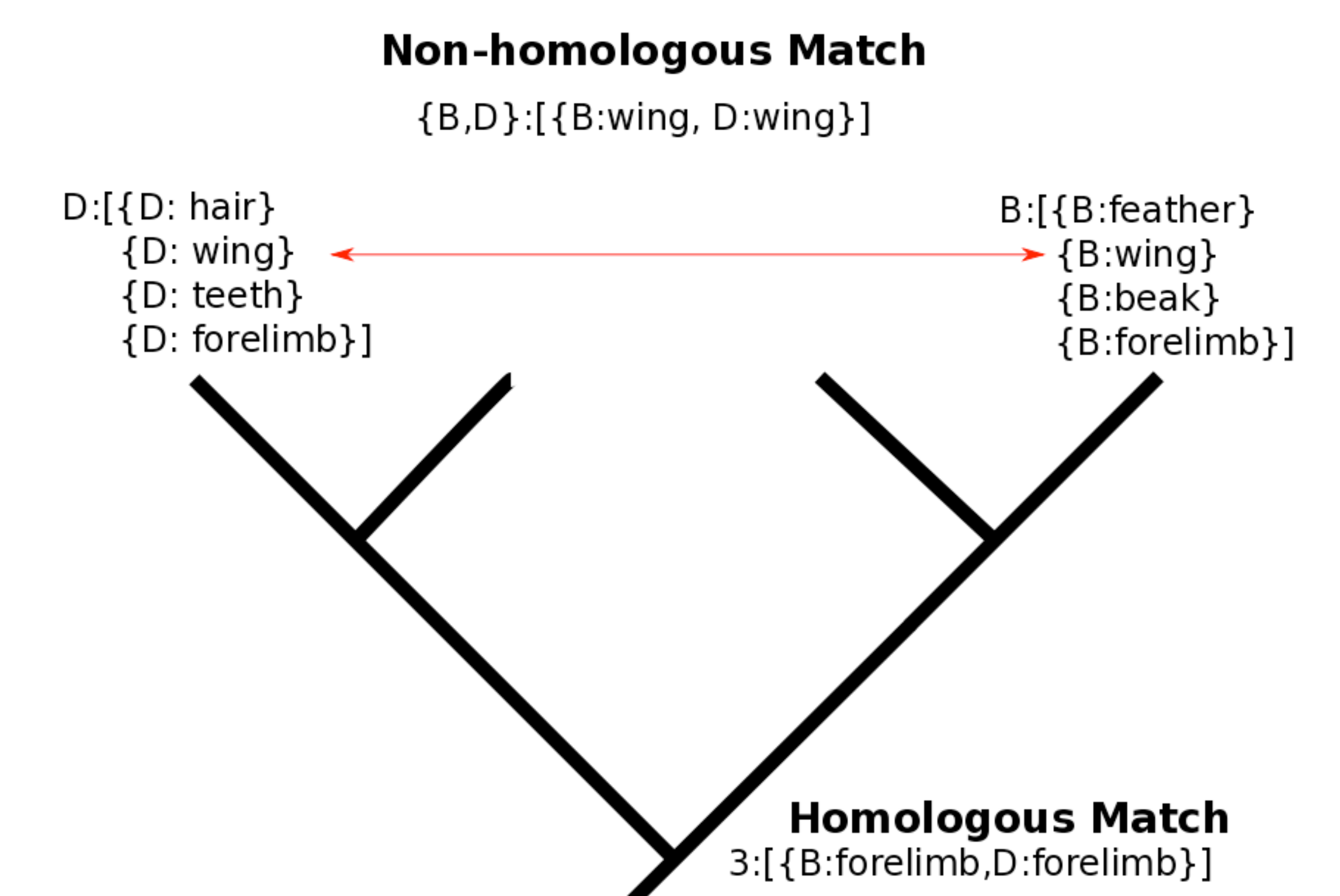
The first step of phylontal is assigning each taxon specific ontology to a tip of the tree. Tips without ontology assignments are ignored. From the assigned ontology, a collection of singleton sets, each containing a term from a filtered subset of the ontology, is constructed and assigned to the corresponding tip.



Traversal begins when by selecting a pair of sibling tips to match. Matching consists of selecting one term from each ontology and constructing a new set containing the terms. The new set is then added to the collection of sets which will be associated with the parent node of the two tips. Matching continues until all matches, either scored above a threshold similarity value or as selected by a user have been made. The result is the collection of two-element sets which is assigned to the parent node.



This is repeated for other pairs of tips with collections of term sets. The procedure for merging interior nodes is similar, except that after the all the matches between sets on the child nodes are complete, the remaining terms on each child node are matched against the children of the sibling ('nieces'). This modification allows polytomies to be resolved arbitrary into bifurcations, however, in such cases, the search for matching terms may require looking further up the descendants terms in the polytomy.



For non-homologous correspondences (e.g., similarity due to convergence), there are also collections of sets of corresponding terms, but as there is no one common ancestor, there is a set of nodes associated with the collection, one for each independent appearance of the similarity in the tree. Such correspondences will be found in a separate search that follows the homology search and would be more exhaustive.

Discussion

This method allows construction of sets of homologous similarities using a fairly simple down-pass over the tree. This has the advantage of reducing the amount of realignment in the case of either uncertainty in a tree or the need to align across multiple trees of similar topology.

Implementation

A library (jar file) and demonstration application implementing this method is under development that will support alignment of taxon specific ontologies in either OWL or OBO and trees in the NeXML⁵ format. A choice of several lexical and structural match methods will be included. Progress of the phylontal library and application can be followed on the project's home at google code (<http://code.google.com/p/phylontal/>).

Acknowledgements

Initial development of these ideas was supported by a short-term visitor fellowship at NESCent during fall 2009, funded under NSF #EF-03423641. Thanks also to the Phenoscope project and to Matthew Bertone for the ongoing discussion of phenotype ontology alignment tools and methods.

References

- Midford, P. E. 2004. Ontologies for Behavior. OUP Bioinformatics 18:3700-3701.
- Euzenat, J and Shvaiko, P. (2007). Ontology Matching Berlin: Springer.
- Bastian, F, Parmentier, G., Roux, J., Moretti, S., Laudet, V., Robinson-Rechavi, M. 2008. Bgee: Integrating and comparing heterogeneous transcriptome data among species. Lecture Notes in Computer Science vol. 5109 pp. 124-131, Berlin: Springer-Verlag.
- http://bgee.unil.ch/bgee/download/homology_ontology_v1.1.obo
- <http://nexml.org/>