OntoPhylo Tutorial: Character Annotation

Diego S. Porto and Sergei Tarasov

23 August, 2023

STEP 0. Install and load packages.

Before starting this tutorial, please verify if you already have *ontophylo* installed. If not, then you can install it by running the code below:

```
remotes::install_github("diegosasso/ontophylo")
```

Now, let's load the package.

```
library(ontophylo)
```

Let's also load some other packages that will be useful in this tutorial. If you have not them installed, please do so by running `install.packages()`.

```
library(ontoFAST)
library(ontologyIndex)
library(tidyverse)
```

STEP 1. Import data.

For this tutorial and the following ones, we will use a modified data set from Sharkey et al. (2012). The modified data set contains 239 morphological characters for 87 species of Hymenoptera.

```
# Import character matrix.
hym_mat <- readRDS("data/hym_matrix.RDS")
```

Overview on character annotation.

Each phylogenetic character is a statement about a particular variable phenotype observed in one or more organisms. In its simplest form, a character statement is an affirmation about the absence/presence or quality of an anatomical entity. More complex statements may relate two or more anatomical entities (e.g. entity A connected to entity B) or include relative measurements (e.g. entity A wider than long).

Character statements are usually represented as strings of text in natural language (=written for humans). In such a format, they cannot be easily parsed or reasoned by computers. A simple alternative to make them computer-parsable is to convert character statements to semantic statements by adopting some structured data model or standardized syntax.

The Entity-Quality (EQ) syntax represents character statements as pairs of linked ontology terms, one from an anatomy ontology and another from a quality ontology. For example, the character statement "Mandible shape: curved" describing the phenotype of a hymenopteran insect can be represented as `Mandible` (HAO_0000506): `curved` (PATO_0000406) by linking an anatomy term from the Hymenoptera Anatomy Ontology (HAO) (Yoder et al., 2010) to a quality term from the Phenotype and Trait Ontology (PATO) (Gkoutos et al., 2005).

For the purposes of the PARAMO pipeline (Tarasov et al., 2019), only the anatomy term is necessary to guide the amalgamation of characters. We just need to link each character statement (i.e. a column in a character matrix) to a specific term from an anatomy ontology that better defines the morphological structure under consideration. This procedure is called 'annotation'. Annotation can be facilitated by using semi-automatic methods such as those implemented in *ontoFAST* (Tarasov et al., 2022) but usually still requires expert-based input to make the final decision. In this tutorial, we will show how to employ pre-annotated characters or semi-automatic annotations to query the ontology.

STEP 2A. Querying the ontology with pre-annotated characters.

Let's start with pre-annotated characters. First, import the table with pre-cooked ontology annotations running the code below:

```
# Import annotations.
hym_annot <- readRDS(file = "data/hym_annot.RDS")
```

To guide the character amalgamations in the PARAMO pipeline, we need first to group characters based on their ontology annotations. In this tutorial, we will use anatomy terms from the Hymenoptera Anatomy Ontology. Terms from this ontology were manually annotated to all character statements, as seen in the `hym_annot` object. In the following chunk of code, we will first import the HAO ontology in OBO format using the package *ontologyIndex* (Greene et al., 2017). Then, we will make a query to retrieve all characters annotated with terms that are *part_of* the three main anatomical regions of the Hymenoptera anatomy: head, mesosoma and metasoma.

```
# Import HAO.
onto <- get_OBO("data/HAO.obo", extract_tags = "everything",
                propagate_relationships = c("BFO:0000050", "is_a"))

# Define query terms for anatomical regions.
query_anat <- c("head", "mesosoma", "metasoma")

# Set a table of matches between characters and ontology terms.
char_info <- hym_annot %>% select(char_id, onto_id)

# Group characters by anatomical regions.
HYM_ANAT <- RAC_query(char_info, onto, query_anat)
```

```
##
## Aggregations by :
## $head
##  [1] "CH1"  "CH2"  "CH3"  "CH4"  "CH5"  "CH6"  "CH7"  "CH8"  "CH10" "CH11"
## [11] "CH12" "CH13" "CH16" "CH21" "CH27" "CH28" "CH29" "CH31" "CH36" "CH37"
## [21] "CH38" "CH39" "CH40" "CH41" "CH42" "CH43" "CH44" "CH45" "CH46" "CH47"
## [31] "CH48" "CH49" "CH50" "CH51" "CH53" "CH54" "CH57" "CH58" "CH59" "CH61"
## [41] "CH62" "CH63" "CH65" "CH66" "CH67" "CH68" "CH71" "CH72" "CH73" "CH74"
```

```
##
## $mesosoma
##   [1] "CH75"  "CH77"  "CH80"  "CH81"  "CH82"  "CH83"  "CH84"  "CH85"  "CH86"
##  [10] "CH87"  "CH88"  "CH89"  "CH90"  "CH95"  "CH97"  "CH99"  "CH100" "CH102"
##  [19] "CH103" "CH104" "CH105" "CH106" "CH109" "CH110" "CH111" "CH112" "CH115"
##  [28] "CH116" "CH117" "CH118" "CH119" "CH120" "CH121" "CH123" "CH124" "CH125"
##  [37] "CH126" "CH127" "CH128" "CH129" "CH130" "CH136" "CH137" "CH141" "CH142"
##  [46] "CH145" "CH146" "CH147" "CH148" "CH149" "CH150" "CH153" "CH154" "CH155"
##  [55] "CH156" "CH157" "CH158" "CH159" "CH160" "CH161" "CH162" "CH163" "CH164"
##  [64] "CH165" "CH169" "CH170" "CH171" "CH172" "CH173" "CH174" "CH175" "CH176"
##  [73] "CH177" "CH178" "CH180" "CH182" "CH183" "CH184" "CH185" "CH186" "CH187"
##  [82] "CH188" "CH189" "CH190" "CH191" "CH192" "CH196" "CH197" "CH198" "CH199"
##  [91] "CH200" "CH201" "CH202" "CH203" "CH204" "CH205" "CH206" "CH209" "CH211"
## [100] "CH214" "CH215" "CH216" "CH217" "CH218" "CH219" "CH220" "CH224" "CH225"
## [109] "CH226" "CH228" "CH229" "CH230" "CH231" "CH233" "CH234" "CH235" "CH236"
## [118] "CH237" "CH240" "CH241" "CH242" "CH243" "CH244" "CH245" "CH246" "CH247"
## [127] "CH248" "CH249" "CH250" "CH251" "CH252" "CH253" "CH254" "CH255" "CH256"
## [136] "CH257" "CH258" "CH261" "CH262" "CH264" "CH265" "CH266" "CH267" "CH268"
## [145] "CH269" "CH270" "CH271" "CH274" "CH353" "CH354" "CH355" "CH356" "CH357"
## [154] "CH358" "CH359" "CH360" "CH363" "CH364" "CH365" "CH366" "CH367" "CH368"
## [163] "CH370" "CH371" "CH372" "CH373" "CH374" "CH378" "CH380" "CH383"
##
## $metasoma
##  [1] "CH279" "CH280" "CH283" "CH284" "CH285" "CH286" "CH287" "CH288" "CH385"
## [10] "CH386" "CH387" "CH389" "CH390" "CH392" "CH393" "CH394" "CH395" "CH396"
## [19] "CH397"
```

STEP 2B. Querying the ontology using *ontoFAST*.

Note that in the previous step, annotations to ontology terms were given in the `onto_id` column.

```
hym_annot %>% slice(1:5)
```

```
## # A tibble: 5 x 4
##   char_id onto_id     term    char
##   <chr>   <chr>       <chr>   <chr>
## 1 CH1     HAO:0000234 cranium Ocellar corona
## 2 CH2     HAO:0000234 cranium Supraantennal groove or depression
## 3 CH3     HAO:0000234 cranium Notch on medial margin of eye
## 4 CH4     HAO:0000234 cranium Position of toruli relative to eyes
## 5 CH5     HAO:0000234 cranium Position of toruli relative to clypeus
```

However, most real-world phylogenetic data sets are not annotated with ontology terms. The only information often available are the character statements written in natural language, for example, as annotations in the character block of a NEXUS file. In our example data set, character statements (not including state descriptions) are shown in the `char` column. Let's try to match some of these with terms from HAO using the *ontoFAST* semi-automatic approach.

```
# Pre-organize the ontology object.
onto$parsed_synonyms <- syn_extract(onto)
onto$id_characters <- hym_annot$char_id
```

```
onto$name_characters <- setNames(hym_annot$char,hym_annot$char_id)

# Run ontoFAST.
auto_annot <- annot_all_chars(onto)
```

```
## [1] "Doing automatic annotation of characters with ontology terms..."
```

Let's check the candidate terms recovered with *ontoFAST* for a small sample of character statements.

```
cand_annot <- setNames(auto_annot[6:8], hym_annot$char[6:8])
lapply(cand_annot, function(x) onto$name[names(onto$name) %in% x] )
```

```
## $`Subantennal shelf`
##                            HAO:0000101                            HAO:0000105
##                              "antenna"                        "antennal shelf"
##                            HAO:0002326
## "maximum diameter of the compound eye"
##
## $`Inner margin of torulus`
##    HAO:0000103    HAO:0000510    HAO:0000908    HAO:0001022    HAO:0001981
## "antennal rim"       "margin"        "scape"      "torulus"       "margin"
##
## $`Subantennal groove`
##          HAO:0000101          HAO:0000965          HAO:0001220
##            "antenna" "subantennal groove" "subantennal groove"
```

As you can see, although *ontoFAST* does a great job in finding potential matches, there is still need for an expert in the taxonomic group to select the terms that best apply to the anatomical entity described in each character statement. Once this is done (we are not going to annotate all +200 characters!), a table similar to `hym_annot` can be constructed and used to query the ontology, as demonstrated before. This table can be saved in CSV format and used in future analyses.

Finally, we can sample 10 characters from each anatomical region for further analyses in the following tutorials and save the results.

```
# Get a sample of characters from each anatomical region.
set.seed(42)
HYM_ANAT <- lapply(HYM_ANAT, function(x) sample(x, 10))
HYM_ANAT
```

```
## $head
##  [1] "CH73" "CH57" "CH1"  "CH42" "CH11" "CH54" "CH31" "CH41" "CH7"  "CH67"
##
## $mesosoma
##  [1] "CH198" "CH372" "CH228" "CH104" "CH358" "CH233" "CH229" "CH252" "CH130"
## [10] "CH383"
##
## $metasoma
##  [1] "CH284" "CH285" "CH390" "CH396" "CH397" "CH280" "CH389" "CH288" "CH283"
## [10] "CH386"
```

```r
# Create a folder to store RData files.
dir.create("RData")

# Save workspace.
save.image("RData/step1_annot.RData")
```

References

Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M., and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome biology*, 6:1–10.

Greene, D., Richardson, S., and Turro, E. (2017). ontologyx: a suite of r packages for working with ontological data. *Bioinformatics*, 33(7):1104–1106.

Sharkey, M. J., Carpenter, J. M., Vilhelmsen, L., Heraty, J., Liljeblad, J., Dowling, A. P., Schulmeister, S., Murray, D., Deans, A. R., Ronquist, F., et al. (2012). Phylogenetic relationships among superfamilies of hymenoptera. *Cladistics*, 28(1):80–112.

Tarasov, S., Mikó, I., and Yoder, M. J. (2022). ontofast: an r package for interactive and semi-automatic annotation of characters with biological ontologies. *Methods in Ecology and Evolution*, 13(2):324–329.

Tarasov, S., Mikó, I., Yoder, M. J., and Uyeda, J. C. (2019). Paramo: A pipeline for reconstructing ancestral anatomies using ontologies and stochastic mapping. *Insect Systematics and Diversity*, 3(6):1.

Yoder, M. J., Mikó, I., Seltmann, K. C., Bertone, M. A., and Deans, A. R. (2010). A gross anatomy ontology for hymenoptera. *PloS one*, 5(12):e15991.