INTRODUCTION

During the past few decades, a massive number of organismal phenotypes have been described by biologists for phylogenetic purposes in the form of character matrices where character statements are written in natural language (NL). Although the matrices can be straightforwardly handled by current phylogenetic methods, a bulk of important biological information included in the NL statements remain unexplored and neglected by the available methods (Deans et al., 2015; Tarasov & Géniér, 2015). This information refers to the complex hierarchical structure of organismal anatomies (e.g. digits are parts of limbs which, in turn, are parts of body) and diverse semantics of anatomical entities (e.g. shape and colouration of digits and limbs).

APPLICATION

ontoFAST: An R package for interactive and semi-automatic annotation of characters with biological ontologies

Sergei Tarasov1,2 | István Mikó3 | Matthew Jon Yoder4

1Finnish Museum of Natural History, Helsinki, Finland
2National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, USA
3University of New Hampshire, Durham, NH, USA
4Illinois Natural History Survey, Champaign, IL, USA

Correspondence
Sergei Tarasov
Email: sergei.tarasov@helsinki.fi

Funding information
National Science Foundation, Grant/Award Number: DBI-1300426; Suomen Akatemia, Grant/Award Number: 331631

Handling Editor: Samantha Price

Abstract

1. Ontologies are becoming a fundamental technology for analysing phenotypic data. The commonly used Entity–Quality (EQ) provides rich semantics for annotating phenotypes and characters using ontologies. However, EQ syntax might be time inefficient if this granularity is unnecessary for downstream analysis.
2. We present an R package ontoFAST that aids fast annotations of characters with biological ontologies. ontoFAST takes a biomedical ontology in OBO format and a list of characters as input, and produces a list of mappings from characters to ontology terms as output.
3. The annotations produced by ontoFAST can be exported in CSV format for downstream analysis. Additionally, ontoFAST provides (a) functions for constructing simple queries of characters against ontologies and (b) helper function for exporting and visualizing complex ontological hierarchies and their relationships.
4. ontoFAST enhances integration of ontological and phylogenetic methods and supports data interoperability between R applications. Ontology tools are underrepresented in R ecosystem and we hope that ontoFAST will stimulate their further development.

KEYWORDS
annotation, character matrix, characters, ontology, phenomics, phenotypes, phylogenetics

1 | INTRODUCTION

Nowadays, ontologies are becoming a fundamental technology for analysing and managing phenotypic data (Balhoff et al., 2010). An ontology is a logic-based representation of concepts and their relationships from a specific domain of knowledge (Balhoff et al., 2013; Deans et al., 2012, 2015). In biosystematics, ontologies have been successfully used for formalizing organismal anatomies (Mungall et al., 2012; Yoder et al., 2010), annotating phenotypes and character matrices (Balhoff et al., 2010), extracting presence/absence traits (Balhoff et al., 2013; Dececchi et al., 2015) and making phenotype-to-genotype predictions (Edmunds et al., 2015; Mabee et al., 2018; Shefchek et al., 2020). Integration of ontologies and phylogenetics is an emerging area of research that promises to unfold the hierarchical and semantic
information encoded in the NL statements for evolutionary inference (Tarasov, 2019). For example, PARAMO pipeline (Tarasov et al., 2019) reconstructs phenotypic evolution at different levels of anatomical hierarchy (or any other biologically reasonable hierarchy) ranging from entire organismal anatomy to its component body regions and treats them as single complex characters. This reconstruction is achieved through ontology-informed amalgamation of individual character histories (i.e. character stochastic maps), which is performed by linking characters with the terms from an anatomy ontology. This way allows viewing the characters not just as an ensemble of character state tokens but as entities that have their own biological meaning.

To be available for ontology-based analyses, NL statements should be annotated with ontology terms. The Entity–Quality (EQ) syntax, adopted by the Open Biological and Biomedical Ontologies Consortium (Washington et al., 2009), and Phenoscape project (http://phenoscape.org), is the common convention for such representation (Balhoff et al., 2010; Gkoutos et al., 2005). EQ syntax binds an entity, corresponding to a specific anatomical structure from an anatomy ontology (e.g. digit or limb) with a quality term (e.g. shape or colouration) from the generic Phenotype and Trait Ontology (PATO, Mungall et al., 2010). The annotation of character matrices and phenotypes using EQ syntax is implemented in the Java-based application Phenex (Balhoff et al., 2010), which provides rich semantics for describing nearly any organismal phenotype at the very high level of granularity (Dahdul et al., 2010, 2018).

However, if the high granularity is not needed in an analysis, then the use of EQ syntax might be time inefficient. For example, PARAMO pipeline requires an annotation file in CSV format to be generated for downstream analyses where a character statement is mapped to one or more ontology term(s) (i.e. IRI(s): internationalized resource identifier). Instead of full EQ information, as produced by tools like Phenex, this CSV file requires only the ‘quality’ or ‘entity’ part of EQ depending on the purpose of an analysis. For example, reconstructing dynamics of ancestral body parts would require linking characters with ‘entity’ terms but comparing evolution of all characters that bear colouration quality versus those that bear shape quality would require linking them with ‘quality’ terms. To date, there is no software that would facilitate this ‘light’ version of phenotypic annotation, at the same time, doing this manually is laborious due to enormous amount of terms contained in any ontology. To fill up this gap, we have created an open-source R (R Core Team, 2020) package ontoFAST that enables ‘light’ phenotypic annotation of characters with biological ontologies.

ontoFAST provides an interactive interface for generating CSV files that map character names to ontology terms, initially through fuzzy matching, and then by manual vetting that can be performed by a scientist familiar with the dataset and ontology being used. Additionally, ontoFAST includes functions for queering characters against ontologies, and visualizing hierarchies of characters and ontologies by means of the sunburstR package (Bostock et al., 2020) and Cytoscape (https://cytoscape.org).

We anticipate that further development of ontology-informed phylogenetic methods would be fuelled if a significant part of applications is kept within the same programming ecosystem. The R ecosystem, unlike its popularity in the phylogenetic community, is not traditionally used for building ontology-oriented packages, which inhibits integration of the ontological and phylogenetic fields (Tarasov, 2019). Currently, the output of ontoFAST can be directly used in PARAMO pipeline and we hope that ontoFAST will help enhancing interoperability between future ontology-informed phylogenetic applications in R.

2 | MATERIALS AND METHODS

2.1 | ontoFAST availability

The current version 1.0.0 of the package ontoFAST requires R 3.5.0 and is distributed under the GPL license. The package can be downloaded from CRAN at https://cran.r-project.org/web/packages/ontoFAST/index.html, its development version is available at https://github.com/sergeitarasov/ontoFAST. The detailed tutorial is given at https://github.com/sergeitarasov/ontoFAST/wiki.

2.2 | Implementation of ontoFAST

ontoFAST uses Shiny (RStudio Inc, 2020) that enables building interactive web applications straight from R. The interactive interface of ontoFAST can be run either from within RStudio (RStudio Team, 2021) or any web browser. ontoFAST uses functions from ontologyIndex package (Greene et al., 2017) for parsing and manipulating ontologies. It also depends on visNetwork package (Almende et al., 2019) for interactive visualization of ontology graphs.

2.3 | Testing

We tested ontoFAST by annotating two character matrices. All these datasets are included in the package and tutorial. One matrix with 392 characters [dataset Sharkey _2011] from the large-scale Hymenoptera (sawflies, wasps, ants and bees) phylogeny (Sharkey et al., 2012) was annotated using the Hymenoptera Anatomy Ontology (dataset HAO) (Yoder et al., 2010); the annotations are stored in Sharkey _2011 _annot dataset.

Another matrix of 232 characters from the dung beetle (Coleoptera: Scarabaeinae) phylogeny (Tarasov, 2017) was annotated using dung beetle ontology (Scarab dataset), the annotations are stored in Tarasov _2017 _annot. The Scarab ontology was developed from HAO by enriching it with anatomical terms specific for dung beetle since beetles, so far, lack any comprehensive anatomy ontology. Scarab is an informal and experimental ontology and should be used with caution in other studies.
3 | RESULTS AND DISCUSSION

3.1 | Input and output data

The character annotation using ontoFAST requires two initial pieces of data: a list of character statements and a biomedical ontology. The list of character statements can be imported into R as a CSV table or a vector of text strings. To import statements from a character matrix stored in the widely used NEXUS format, one can open it in a popular software Mesquite (Maddison & Maddison, 2018) and copy character statements into any software that supports CSV format (e.g. Microsoft Excel or Atom https://atom.io/).

Any organism-specific anatomy ontology or supporting ontologies (e.g. PATO, BSPO, Dahdul et al., 2014, RO Mungall et al., 2021) can be used for annotating characters in ontoFAST; the selected ontology should be in OBO format and can be read with get_OBO() function from ontologyIndex package.

```r
install.packages("ontology")
install.packages("igraph")
library("ontology")
```

The output of ontoFAST is an R list object that contains annotations: character IDs and associated IRIs (or names) of ontology terms. This list can be used for queries, exported to CSV format, or, using special functions provided by ontoFAST, exported to third-party applications.

3.2 | Annotating characters using ontoFAST

Prior to running the interactive interface, the read-in data have to be preprocessed in R console using the following three steps. First, run onto_process() function to combine ontology and character statement into a single object of ontology-index class (Greene et al., 2017); this function automatically parses synonyms from the ontology, the argument do.annot = TRUE runs fuzzy matching of characters against the ontology and suggests candidate terms for annotation. Second, create a new environment to store a global variable that will serve as an input and output for the interactive mode; the separate environment is required due to CRAN policy restricting the usage of global variables (https://cran.r-project.org/web/packages/policies.html); it should be specifically called ontofast, other names will not work due to the way R handles environments. Third, use the function make_shiny_in() to create an object in the ontofast environment; the name of this variable is taken as an argument by runOntoFast() to launch the interactive session.

```r
ontofast <- new.env(parent = emptyenv())
# creating shiny_in variable to serve as an input and output for runOntoFast()
ontofast$shiny_in <- make_shiny_in(hao_obo)
# running the interactive session
runOntoFast(is_a = c("is_a"), part_of = c("BFO:0000050"), shiny_in="shiny_in",
file2save = "ontoFAST_shiny_in.RData")
```

The interactive interface consists of four panels (Figure 1). The ontology panel shows a graph where nodes are the ontology classes and edges are part_of and is_a relationships, other relationships are not yet supported. The customize panel on the top of the window allows selecting relationships to display and navigate to a required term by typing in a few first letters of its name. The information panel shows ID, synonyms and definition of the term selected in the navigation panel.

The leftmost character panel shows the character statements. There are three ways to annotate them: (a) if you ran fuzzy matching with onto_process(), the candidate terms are shown below the ‘Add’ button and can be selected by checking the respective box(es); (b) click on a node in the ontology panel, move the cursor to the character panel and click the ‘Add’ button under the respective character statement; and (c) paste term IRI right in the character panel. Every character can be annotated with more than one term.

Upon the annotation is complete, you can close the window and return to the console mode. The characters and their annotation are stored in the lists ontofast$shiny_in/terms_selected and ontofast$shiny_in/terms_selected_id. Consider saving your data, by clicking the ‘Save file’ button in the top right corner while in the interactive mode. This will help to avoid risk of losing annotations, if an R session crashes; ontoFAST saves data to the specified file via file2save argument in runOntoFast(). The created annotations can be further used in downstream analyses or saved as CSV file.

```r
write.csv(out, "annotations.csv")
```

3.3 | Efficiency of ontoFAST and comparison to other annotation services

The ‘light’ approach for characters annotation—tagging characters with selected ontology terms—offered by ontoFAST begins with the automatic search of the candidate terms using the fuzzy matching. The suggested terms should be further selected manually, which requires familiarity with an ontology being used. Since ontologies evolve constantly, even a familiar user may experience difficulty...
with selecting the right terms. To facilitate it, ontoFAST provides an interactive interface for graphical navigation through the terms and their relationships. If the ‘light’ annotation is appropriate for an analysis, these features help reducing the annotation time in comparison to the traditional EQ approach.

Other service can be potentially used for annotating characters with ontologies. For example, The Monarch Initiative (Shefchek et al., 2020) offers an integrative platform for connecting phenotypes to genotypes while its SciGraph instance (https://scigraph-ontology.monarchinitiative.org/scigraph/docs/) can annotate free text using powerful algorithms that are superior to the fuzzy search of ontoFAST. This annotation service, focused on human phenotypic data, is not specifically tailored to work with character statements and lacking an interactive interface for manual vetting. However, with additional programming, it might be suitable for annotating characters of Vertebrata using the UBERON ontology (Mungall et al., 2012); its application to other taxa would be challenging at the moment as the Monarch Initiative does not cover all anatomical ontologies.

3.4 | Visualizing and queering annotations

3.4.1 | Visualizing with sunburstR and Cytoscape

Having characters linked with an ontology may provide insights into the anatomical and functional spread across biological entities. We use the annotations produced with ontoFAST for Hymenoptera and dung beetles (see the Data section) to demonstrate it. The hierarchical structure can be visualized using a sunburst plot from the sunburstR package (Bostock et al., 2020). This plot shows relational hierarchy using a series of rings; each ring corresponds to a level in the ontological hierarchy—the inner circles represent ontology classes and outermost circle represents the annotated characters (Figure 2b–d). The function paths.sunburst() automatically converts ontoFAST data to sunburstR format. For example, both Scarabaeinae and Hymenoptera phylogenies include characters from various body regions but in both datasets the thoracic characters (~30%) are predominate (Figure 2b–d).

The annotations and ontologies can be also exported to Cytoscape using export.cytoscape() function for further manipulation and visualization. Cytoscape is an open-source software for visualizing complex networks. The graph obtained with Cytoscape might be used to analyse distribution of characters across ontology classes (i.e. the graph nodes). For instance, the Hymenoptera characters are more or less uniformly distributed across the body regions (i.e. the graph nodes), while those of Scarabaeinae are clustered into groups (Figure 2a,b), which indicates that only the specific body regions have undergone changes in Scarabaeinae during their evolutionary course.

3.4.2 | Querying

Our package has a set of functions for running simple queries with the annotated characters. The function chars.per.term() calculates the number of characters for each ontology class; get_ancestors_chars() return all shared ancestral
ontology classes for a set of characters; and get_descendants_chars() returns all characters descending from a given ontology class.

ACKNOWLEDGEMENT
We are thankful to the two anonymous reviewers and editors for their comments which helped us improve the manuscript.

CONFLICT OF INTEREST
Authors declare no conflict of interest.

AUTHORS’ CONTRIBUTIONS
S.T. conceived and designed the package; I.M. and S.T. annotated Hymenoptera and Scarabaeinae datasets. All authors contributed to the package design and wrote the paper.

FIGURE 2 Visualization of characters annotated using ontoFAST for Hymenoptera and Scarabaeinae (dung beetles). (a, b) The network of ontology classes and the linked characters produced using Cytoscape. (b–d) The plots show hierarchy of characters and ontological classes produced using the sunburstR package

PEER REVIEW
The peer review history for this article is available at https://pubons.com/publon/10.1111/2041-210X.13753.

DATA AVAILABILITY STATEMENT
The code of ontoFAST version 1.0.0 including all data is archived on Zenodo (Tarasov, 2021).

ORCID
Sergei Tarasov https://orcid.org/0000-0001-5237-2330

REFERENCES
Almende, B. V., Thieurmel, B., & Robert, T. (2019). visNetwork: Network Visualization using ‘vis.js’ Library. R package version 2.0.9. https://CRAN.R-project.org/package=visNetwork
Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., & Vision, T. J. (2010). Phenex: Ontological annotation of phenotypic diversity. PLoS One, 5, e105000.

Balhoff, J. P., Mikó, I., Yoder, M. J., Mullins, P. L., & Deans, A. R. (2013). A semantic model for species description applied to the ensign wasps (hymenoptera: Eulophidae) of new caledonia. Systematic Biology, 62, 639–659. https://doi.org/10.1093/sysbio/syt028

Bostock, M., Rodden, K., Warne, K., & Russell, K. (2020). sunburstR: Sunburst ‘Htmlwidget’. https://CRAN.R-project.org/package=sunburstR. R package version 2.1.5

Dahdul, W. M., Balhoff, J. P., Engeman, J., Grande, T., Hilton, E. J., Kothari, C., Lapp, H., Lundberg, J. G., Midford, P. E., Vision, T. J., Westerfield, M., & Mabee, P. M. (2010). Evolutionary characters, phenotypes and ontologies: Curating data from the systematic biology literature. PLoS One, 5, e10708. https://doi.org/10.1371/journal.pone.0010708

Dahdul, W. M., Cui, H., Mabee, P. M., Mungall, C. J., Osumi-Sutherland, D., Walls, R. L., & Haendel, M. A. (2014). Nose to tail: roots to shoots: Spatial descriptors for phenotypic diversity in the biological spatial ontology. Journal of Biomedical Semantics, 5, 1–13. https://doi.org/10.1186/2041-1480-5-34

Dahdul, W., Manda, P., Cui, H., Balhoff, J. P., Dececchi, T. A., Ibrahim, N., Lapp, H., Vision, T., & Mabee, P. M. (2018). Annotation of phenotypes using ontologies: A gold standard for the training and evaluation of natural language processing systems. Database, 2018. https://doi.org/10.1093/database/bay110

Deans, A. R., Lewis, S. E., Hula, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtoy, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., ... Mabee, P. (2015). Finding our way through phenotypes. PLoS Biology, 13, e1002033. https://doi.org/10.1371/journal.pbio.1002033

Deans, A. R., Mikó, I., Wipfler, B., & Friedrich, F. (2012). Evolutionary phenomena and the emerging Enlightenment of arthropod systems. Invertebrate Systematics, 26, 323–330. https://doi.org/10.1071/IS12063

Dececchi, T. A., Balhoff, J. P., Lapp, H., & Mabee, P. M. (2015). Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenomena across studies. Systematic Biology, 64, 936–952. https://doi.org/10.1093/sysbio/syv031

Edmunds, R. C., Su, B., Balhoff, J. P., Eames, B. F., Dahdul, W. M., Lapp, H., Lundberg, J. G., Vision, T. J., Dunham, R. A., Mabee, P. M., & Westerfield, M. (2015). Phenoscape: Identifying candidate genes for evolutionary phenotypes. Molecular Biology and Evolution, 33, 13–24. https://doi.org/10.1093/molbev/msv223

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

Greene, D., Richardson, S., & Turro, E. (2017). ontologyx: A suite of R packages for working with ontological data. Bioinformatics, 33, 1104–1106. https://doi.org/10.1093/bioinformatics/btw763

Mabee, P. M., Dahdul, W. M., Balhoff, J. P., Lapp, H., Manda, P., Uyeda, J., Vision, T., & Westerfield, M. (2018). Phenoscape: Semantic analysis of organismal traits and genes yields insights in evolutionary biology.

Maddison, W., & Maddison, D. (2018). Mesquite: A modular system for evolutionary analysis. version 3.40. https://www.mesquiteproject.org

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., & Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. Genome Biology, 11, 1–16. https://doi.org/10.1186/gb-2010-11-1-r2

Mungall, C., Osumi-Sutherland, D., gpaudet, Overton, J. A., Matentzoglu, N., Clare72, Balhoff, J., Harris, N., Brush, M., Touré, V., Sinclair, M., Poelen, J., Bretaudeau, A., Cain, S., Haendel, M., Vasilevsky, N., diatomsRcool, Hammock, J., Laporte, M.-A., ... Larralde, M. (2021). oborel/obo-relations: 2021-03-08 release. https://doi.org/10.5281/zenodo.4589530

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. Genome Biology, 13, 1–20. https://doi.org/10.1186/gb-2012-13-1-r5

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org

RStudio Inc. (2020). shiny: Easy web applications in R. http://shiny.rstudio.com

RStudio Team. (2021). RStudio: Integrated Development Environment for R. RStudio. http://www.rstudio.com/

Sharkey, M. J., Carpenter, J. M., Vilhelmsen, L., Heraty, J., Liijeblad, J., Dowling, A. P. G., Schulmeister, S., Murray, D., Deans, A. R., Ronquist, F., Krogmann, L., & Wheeler, W. C. (2012). Phylogenetic relationships among superfamilies of hymenoptera. Cladistics, 28, 80–112. https://doi.org/10.1111/j.1096-0031.2011.00366.x

Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X. A., Balhoff, J. P., Babb, L., Bello, S. M., Blau, H., Bradford, Y., Carbon, S., Carmody, L., Chan, L. E., Cipriani, V., ... Osumi-Sutherland, D. (2020). The monarch initiative in 2019: An integrative data and analytical platform connecting phenotypes to genotypes across species. Nucleic Acids Research, 48, D704–D715. https://doi.org/10.1093/nar/gkz997

Tarasov, S. (2017). A cybertaxonomic revision of the new dung beetle tribe parachorii (coleoptera: Scarabaeidae: Scarabaeinae) and its phylogenetic assessment using molecular and morphological data. Zootaxa, 4329, 101–149. https://doi.org/10.11646/zootaxa.4329.2.1

Tarasov, S. (2019). Integration of anatomy ontologies and Evo-Devo using structured Markov models suggests a new framework for modeling discrete phenotypic traits. Systematic Biology, 68(5), 698–716. https://doi.org/10.1093/sysbio/syz005

Tarasov, S. (2021). ontoFAST: An R package for interactive and semi-automatic annotation of characters with biological ontologies. Zenodo. https://doi.org/10.5281/zenodo.5570524

Tarasov, S., & Génier, F. (2015). Innovative bayesian and parsimony phylogeny of dung beetles (Coleoptera, Scarabaeidae, Scarabaeinae) enhanced by ontology-based partitioning of morphological characters. PLoS One, 10, e0116671. https://doi.org/10.1371/journal.pone.0116671

Tarasov, S., Mikó, I., Yoder, M. J., & Uyeda, J. C. (2019). Paramo: A pipeline for reconstructing ancestral anatomies using ontologies and stochastically mapping. Insect Systematics and Diversity, 3, 1. https://doi.org/10.1093/isd/ixz009

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., & Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology, 7, e1000247. https://doi.org/10.1371/journal.pbio.1000247

Yoder, M. J., Mikó, I., Seltmann, K. C., Bertone, M. A., & Deans, A. R. (2010). A gross anatomy ontology for hymenoptera. PLoS One, 5, e15991. https://doi.org/10.1371/journal.pone.0015991

How to cite this article: Tarasov, S., Mikó, I., & 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, 324–329. https://doi.org/10.1111/2041-210X.13753