PhyloSophos

Min Hyung Cho

March 15th, 2023

[1] Summary

PhyloSophos is a high-throughput scientific name processor which achieves greater mapping performance by referencing multiple taxonomic references and recognizing the semantic structure of scientific names. It also corrects common Latin variants and vernacular names, which often appear in various biological databases and resources.

Available at: https://github.com/mhcho4096/phylosophos

Also refer to: (journal reference will be included here)

[2] Installation guide

PhyloSophos requires Numpy, along with other basic libraries of Python 3. Before initializing PhyloSophos to fit with your environment, please set up an environment with a python version >=3.7 (preferably with conda), and install numpy (https://numpy.org/install/) as appropriate.

A script for initialization and update is phylosophos_initialize_update.py. This script will download raw taxonomic metadata from Catalogue of Life (CoL), Encyclopedia of Life (EoL), and NCBI Taxonomy, and then process it into reference files. You may execute this script as:

```
> phylosophos_initialize_update.py [optional_update_parameter]
```

For initialization purposes, the optional parameter will not affect downstream processes. It is possible for Catalogue of Life and Encyclopedia of Life to change the name of the metadata file in their FTP server. If this happens, please change file names in the script (see lines 76 & 77) as appropriate.

In addition, there was an instance when the taxonomic databases changed the format of their metadata, making the initialization script non-functional. If this happens, please notify me immediately (see contact information).

[3] Update guide

Again, taxonomic reference update uses phylosophos_initialize_update.py script. You may execute this script as:

```
> phylosophos_initialize_update.py [optional_update_parameter]
```

In this case, optional update parameter affects the downstream process: by default, if a database metadata file is found in the /external files directory, the metadata download step is skipped to reduce processing time (allowing manual download of reference metadata). If an integer value other than 0 is given as an optional argument, the update script will start downloading the reference metadata file, overriding any pre-existing data.

[4] Usage guide

PhyloSophos analysis could be performed by executing phylosophos_core.py script. You may execute this script as:

```
> phylosophos_core.py [[optional_parameter_type] [optional_parameter_value]]
```

PhyloSophos currently recognizes five types of optional parameter types.

- Help (-h, -help, -guide): if one of these arguments is given, a hard-coded guide to PhyloSophos will appear in the console. This will provide simple instructions on how to customize PhyloSophos mapping parameters. No following parameter value is required.
- Reference type change (-r, -ref): if one of these arguments is given, PhyloSophos will change the database of choice to the one specified by the following argument. The default setting is 'ncbi', while 'col' and 'eol' are also available in basic PhyloSophos system. You may change the default setting by modifying /ps_init/ps_initialize.py (see lines 56, 60 & 62). If you want to include other types of references into PhyloSophos system, please read chapter 6.
- Input type change (-i, -input): if one of these arguments is given, along with the name of the input file, PhyloSophos will specifically import the given file as an input. If not (as a default setting), PhyloSophos will consider all files within the /input directory to be scientific name input files.
- Levenshtein distance cutoff (-l, -lev, -cutoff): if one of these arguments is given, along with an integer value, PhyloSophos will change the edit distance cutoff (default setting = 3) to the specified value.
- Manual curation status (-m, -manual, -curation): if one of these arguments is given, along with a value 1, PhyloSophos will import /pp_learning/manual_curation_list.tsv and utilize this information to pre-process inputs. If not (as a default setting), PhyloSophos will not import extra information other than reference data files within /pp_ref directory.

[5] Result format guide

The results of the PhyloSophos analysis will be deposited in the /result directory. The name of the result file will be 'phylosophos_result_[export_date]_[export_time]_[input_file_name]'. [export_date] and [export_time] will be six-digit numbers.

PhyloSophos result file has the following format (delimited by tabs).

[Input_file_name] - [Input_original_order] - [Raw_name_input] - [Pre_corrected_input] - [Chosen_reference] - [Chosen_reference_mapped_ID] - [Chosen_reference_scientific_name] - [Chosen_reference_mapping_status_code] - [Chosen_reference_mapping_status_description] - [[(specific_reference)_mapped_id] - [(specific_reference)_scientific_name] - [(specific_reference)_mapping_status_code]] - [Manual_curation_recommended]

Each column shows the information about:

- [Input_file_name]: Name of input file which contains given scientific name input

- [Input_original_order]: The order in which the scientific name is contained in the file

- [Raw_name_input]: Raw scientific name input, as it is appeared in the input file

- [Pre_corrected_input]: Pre-corrected scientific name input

- [Chosen_reference]: Reference of choice (e.g. 'ncbi', 'col', 'eol')

- [Chosen_reference_mapped_ID]: Taxonomic entry ID(s) mapped to given scientific name input (based on reference of choice)

- [Chosen_reference_scientific_name]: Canonical scientific name(s) associated with mapped taxonomic ID(s) (based on reference of choice)

- [Chosen_reference_mapping_status_code]: PhyloSophos mapping status code (see chapter 7) (based on reference of choice)

- [Chosen_reference_mapping_status_description]: Short description of mapping status code (see chapter 7) (based on reference of choice)

- [(specific_reference)_mapped_id]: Taxonomic entry ID(s) mapped to given scientific name input (based on specific reference)

- [(specific_reference)_scientific_name]: Canonical scientific name(s) associated with mapped taxonomic ID(s) (based on specific reference)

- [(specific_reference)_mapping_status_code]: PhyloSophos mapping status code (see chapter 7) (based on specific reference)

- [Manual_curation_recommended]: Mapping status code-based opinion on whether manual curation is needed for this mapping result

Each taxonomic reference found within /pp_ref directory provides [(specific_reference)_mapped_id] - [(specific_reference)_scientific_name] - [(specific_reference)_mapping_status_code] column triplet. Base PhyloSophos provides 3 column triplets for CoL/EoL/NCBI taxonomy respectively.

[6] Modding guide: inclusion of additional taxonomic reference

PhyloSophos extracts two files from each taxonomic database and uses them as references. Each file has the following format (delimited by tabs)

- [ref]_node_dict.txt: [reference_identifier] - [reference_entity_canonical_name] - [reference_entity_synonyms] - [reference_entity_taxonomic_rank] - [reference_entity_phylogenetic_lineage] - [reference_entity_lineage_taxonomic_rank]

  - reference_identifier: accession numbers of given entity within a database.

  - reference_entity_canonical_name: canonical scientific name which the taxonomic database suggests for the given taxonomic entity.

  - reference_entity_synonyms: alternative names which the taxonomic database provides for the given taxonomic entity. The list is represented as a string, which is divided with '|' character.

  - reference_entity_taxonomic_rank: taxonomic rank of given taxonomic entity.

- reference_entity_phylogenetic_lineage: full taxonomic lineage of given taxonomic entity. Starting from the entity itself on the far left, the parent entity is successively given to the right, ultimately leading to the root of the phylogenetic tree. The list is represented as a string, which is divided with '|' character.

- reference_entity_lineage_taxonomic_rank: taxonomic ranks for each entities within a phylogenetic lineage. 'canonical' taxonomic ranks are given a number greater than 0, while all other ranks are given a number 0. The list is represented as a string, which is divided with '|' character.

  - 1: domain, 2: kingdom, 3: phylum, 4: class, 5: order, 6: family, 7: genus, 8: species

● [ref]_genus_dict.txt: [internal_order] - [first_word_block] - [associated_reference_identifier_list]

- internal_order: numerical order of word-blocks (not necessary).

- first_word_block: all word-blocks found within scientific names (canonical names & synonyms) found within a database.

- associated_reference_identifier_list: list of all accession numbers (see reference identifier) which have a preceding word appearing as the first word in one of the associated names. The list is represented as a string, which is divided with '|' character.

If reference metadata is processed into the specified formats and added to the /pp_ref directory, PhyloSophos will function normally. gbif_preprocessing.py is an example script which downloads GBIF backbone taxonomy to /external_files directory, processes Taxon.tsv file to gbif_node_dict.txt and gbif_genus_dict.txt, incorporating GBIF into local PhyloSophos system.

On the other hand, it is possible to remove a particular reference from the /pp_ref directory, reducing the number of references PhyloSophos refers to. While doing so reduces the reference import time, it does not significantly affect processing time itself, as it increases the number of input strings to be corrected (status codes >= 20). This also increases the number of 'incorrectly corrected' inputs, such as correcting valid scientific names and mapping them to incorrect DB entries. So we highly discourage to remove taxonomic references from the /pp_ref directory.

[7] PhyloSophos modding guide: incorporation of manual curation data

There is 'manual_curation_list.tsv' file within the /pp_learning directory. If the manual curation option is activated, PhyloSophos imports this file and uses it to convert obsolete synonyms and vernacular name inputs.

Manual curation file has two columns: 'Common_name' column and 'Curated_name' column. The 'Common_name' column contains names that need to be corrected, while the 'Curated_name' column contains the corresponding scientific names that have been manually identified.

You could provide as many [name] - [changed name] pairs as possible, as long as there are no duplicate entries in the 'common_name' column. If there is a duplicate entry, the previous pair of information will be ignored.

If the curated name is found to be non-biological (e.g. minerals used in traditional medicine), please enclose the changed name with '<<' and '>>' brackets. PhyloSophos will recognize these flags and return a non-organism flag (mapping status code 90) instead of applying the name input into futile edit-distance mapping steps, thus reducing the processing time.

[8] Mapping status code description

Detailed descriptions of each mapping status codes are as follows.

● Codes 0~9: Valid scientific name/Exact match.

- Code 0: Exact match found within a DB of choice / No correction / matched with a canonical name

  - Input is found within a DB of choice as-is. It is matched with canonical name of a taxonomic entity.

- Code 1: Exact match found within a DB of choice / No correction / matched with a single synonym

  - Input is found within a DB of choice as-is. It is matched with synonym of a taxonomic entity.

- Code 2: Exact match found within a DB of choice / No correction / matched with multiple entities

  - Input is found within a DB of choice as-is. It is matched with multiple taxonomic entities.

  - It is possible that a single scientific name, without proper authority information, could refer to multiple entities within a taxonomic database. Most of these cases involve single-word input and fall into one of three categories:

    - Genus and subgenus: Many subgenera share the same epithet as the genera they are included in (e.g. subgenus Rhododendron within genus Rhododendron).

    - Taxa in different phylogenetic domains: Scientific names are governed by multiple nomenclature codes, which govern specific types of taxonomic groups (e.g. animals, plants, bacteria). As these nomenclature codes are independent of one another, it is possible that a single generic epithet is used for multiple genera in different taxonomic groups (e.g. Anisoptera, Callicarpa, Darwinella).

    - Obsolete synonyms: There were several instances where a single scientific name was applied to multiple taxa (e.g. Polyporus badius). As it is a violation of the nomenclature code, misapplied names are later corrected by taxonomic authorities; however, those names remain in a record of obsolete synonyms. If a given input name is not found in a canonical list of scientific names but is found multiple times in a synonym list, PhyloSophos will return all entities associated with that name.

  - Manual review is highly recommended in these cases.

- Code 3: Exact match found within a DB of choice / Simple correction / matched with a canonical name

  - Input is found within a DB of choice after a simple correction process. (similar to code 0) It is matched with canonical name of a taxonomic entity.

- Code 4: Exact match found within a DB of choice / Simple correction / matched with a single synonym

  - Input is found within a DB of choice after a simple correction process. (similar to code 1) It is matched with synonym of a taxonomic entity.

- Code 5: Exact match found within a DB of choice / Simple correction / matched with multiple entities

  - Input is found within a DB of choice after a simple correction process. (similar to code 2) It is matched with multiple taxonomic entities.

- Code 6: Exact match found within other DBs / Synonym information is used to search in a DB of choice / Matched with a single entity

  - Input is found in more than one of the other taxonomic databases, but not in the DB of choice. To identify the corresponding taxonomic entity within a DB of choice, we collected synonym information from other databases and searched for these names again in a DB of choice. As a result, (similar to codes 0,1) a single taxonomic entity was identified.

- Code 8: Exact match found within other DBs / Synonym information is used to search in a DB of choice / Matched with multiple entities

    - Input is found in more than one of the other taxonomic databases, but not in the DB of choice. To identify the corresponding taxonomic entity within a DB of choice, we collected synonym information from other databases and searched for these names again in a DB of choice. As a result, (similar to code 2) multiple taxonomic entities were identified.

● Codes 10~19: Valid scientific name / Nearest taxon match.

- If a given input is mapped to a taxonomic entity with a mapping code of less than 10, PhyloSophos considers it to be a valid scientific name and does not consider it a target for edit distance-based correction. If a given input is not found within a DB of choice, nearest taxon match algorithm is applied instead.

- PhyloSophos extracts phylogenetic lineage information from a taxonomic entity that corresponds to a given input. Starting from the lowest taxonomic rank, the algorithm searches for the taxonomic entity within a DB of choice that exactly matches (as code 0) the name of the higher-rank taxon. This method allows to identify the lowest ('nearest') taxonomic entity which includes the given input as a member.

- Different codes are applied based on the taxonomic level of the matched taxon.

    - Code 10: Nearest match to the species level

    - Code 11: Nearest match to the genus level

    - Code 12: Nearest match to the family level

    - Code 13: Nearest match to the order level

    - Code 14: Nearest match to the class level

    - Code 15: Nearest match to the phylum level

    - Code 16: Nearest match to the kingdom level

    - Code 17: Nearest match to the domain level

● Codes 20~29: Specific epithet corrected.

- Code 20: Specific epithet corrected / Exact match found within a DB of choice / Matched with a single entity

    - Corrected input is found within a DB of choice. It is matched with a single taxonomic entity within a DB.

- Code 21: Specific epithet corrected / Exact match found within a DB of choice / Matched with multiple entities

    - Corrected input is found within a DB of choice. It is matched with multiple taxonomic entities within a DB.

- Code 22: Specific epithet corrected / Exact match found within other DBs / Synonym information is used to search in a DB of choice / Matched with a single entity

    - Corrected input is found in more than one of the other taxonomic databases, but not in the DB of choice. To identify the corresponding taxonomic entity within a DB of choice, we collected synonym information from other databases and searched for these names again in a DB of choice. As a result, a single taxonomic entity was identified.

- Code 23: Specific epithet corrected / Exact match found within other DBs / Synonym information is used to search in a DB of choice / Matched with multiple entities

- Corrected input is found in more than one of the other taxonomic databases, but not in the DB of choice. To identify the corresponding taxonomic entity within a DB of choice, we collected synonym information from other databases and searched for these names again in a DB of choice. As a result, multiple taxonomic entities were identified.

- Code 24: Specific epithet corrected / Exact match found within other DBs / Nearest taxon match

- If a corrected input is mapped to a taxonomic entity with a mapping code of less than 24 in one of the other databases, but not in the DB of choice, the nearest taxon match algorithm is applied (see codes 10~19).

● Codes 30~39: Generic/specific epithet corrected.

- Code 30: Latin inflection corrected / Exact match found within a DB of choice

- One of the possible 'original forms' of a given input is found within a DB of choice.

- Code 31: Latin inflection corrected / Exact match found within other DBs

- One of the possible 'original forms' of a given input is found in more than one of the other taxonomic databases, but not in the DB of choice. This code applies to both synonym-match (as code 3) and nearest taxon match (as code 24).

- Code 32: Generic/Specific epithet corrected / Exact match found within a DB of choice / Matched with a single entity

- Corrected input is found within a DB of choice. It is matched with a single taxonomic entity within a DB.

- Code 33: Generic/Specific epithet corrected / Exact match found within a DB of choice / Matched with multiple entities

- Corrected input is found within a DB of choice. It is matched with multiple taxonomic entities within a DB.

- Code 34: Generic/Specific epithet corrected / Exact match found within other DBs / Synonym information is used to search in a DB of choice / Matched with a single entity

- Corrected input is found in more than one of the other taxonomic databases, but not in the DB of choice. To identify the corresponding taxonomic entity within a DB of choice, we collected synonym information from other databases and searched for these names again in a DB of choice. As a result, a single taxonomic entity was identified.

- Code 35: Generic/Specific epithet corrected / Exact match found within other DBs / Synonym information is used to search in a DB of choice / Matched with multiple entities

- Corrected input is found in more than one of the other taxonomic databases, but not in the DB of choice. To identify the corresponding taxonomic entity within a DB of choice, we collected synonym information from other databases and searched for these names again in a DB of choice. As a result, multiple taxonomic entities were identified.

- Code 36: Generic/Specific epithet corrected / Exact match found within other DBs / Nearest taxon match

- If a corrected input is mapped to a taxonomic entity with a mapping code of less than 36 in one of the other databases, but not in the DB of choice, the nearest taxon match algorithm is applied (see codes 10~19).

● Codes 40~49: Correction denied.

- Code 40: Strain name involved / Nearest match.

- Strain codes are often very similar to one another, but they do not reflect the phylogenetic relationship between individual strains. Therefore, edit distance-based correction approaches could suggest scientific names with strain information, which are almost identical string-wise but not accurate in a taxonomic sense. If word-blocks containing strain-specific information (usually numeric characters) are identified within an input string, PhyloSophos removes them from the correction process and maps them to the nearest higher taxon (e.g. species).

- Code 41: Similarity-related abbreviation identified / Nearest match.

- Abbreviations such as "cf.", "aff.", and "sp.nov." denote similarity to the previously described taxon, but not necessarily identity. Therefore, it may be inappropriate to drop this abbreviation and revert to the scientific name of a taxon that is similar to the species which the input name describes. Instead, PhyloSophos first removes the abbreviations first, tries to match it, and then returns an identifier of a higher taxon which is found within the phylogenetic lineage of the match, to reflect the connotation of given abbreviation.

● Codes 90~99: Mapping denied.

- If a given input is not found within one of the databases as-is, and if it contains word-blocks that require special attention, then the further mapping process is denied and one of these codes is returned.

- Code 90: Non-organism flags

- Code 91: Unclassified-Uncultured-Unidentified

- Code 92: Environmental samples

- Code 93: Virus or phage

- Code 94: Phytoplasma

- Code 95: (endo)symbiont

- Code 96: Unresolvable hybrid

- Code 97: Multiple materia medica

- Manual review is highly recommended in these cases.

● Code 100: Partial mapping.

- If taxonomic mapping is not possible with a full input string, PhyloSophos attempts to find a taxonomic entity which matches a part of the input. It consecutively removes a word-block from the right side of the input string and searches for an exact match: if an exact match is found, the search process is terminated and a result is returned.

- It usually ends with genus or species level mapping (but not always).

● Code 1000: Unmapped input.

- For inputs that cannot be assigned to a corresponding taxonomic entity (not even partially), they are given the code.

[9] Contact information

If you have any questions or issues with PhyloSophos, please contact me: mhcho@bmdrc.org