Taxonomic name resolution using the Leipzig Plant Catalogue and `lcplants`

Alexander Zizka

Background

When comparing or merging any list of plant species, for instance for compiling regional species list, merging occurrence lists with trait data or phylogenies the taxonomic names must be matched, to avoid artificial inflation due to synonyms, data loss or erroneous matches due to homonyms.

The Leipzig Plant Catalogue

The Leipzig Plant Catalogue is a novel global taxonomic backbone, updating The Plant List, comprising more than 1,300,000 names and 350,000 accepted names. We described the LCP in detail in the related scientific publication (NOT YET PUBLISHED).

lcplants

A package for large-scale taxonomic harmonization of plant names by fuzzy matching and synonymy resolution against the Leipzig Plant Catalogue as taxonomic backbone. Submission of single names or list of species names is possible (for lists with more than 5000 species computation may take some time).

Running lcplants

Installation

You can install `lcplants` from github using the devtools package (you may need to install devtools first). To use `lcplants` you also need the data of the `LCP` package, which you can install in the same way.

```
library(devtools)
devtools::install_github("idiv-biodiversity/lcplants")
devtools::install_github("idiv-biodiversity/LCP")
library(lcplants)
```

Input

All you need to run lcplants is a species name or a list of species names for which you need the status or synonymy. The genus name and epitheton should not contain special characters, the name of the authorities might contain special characters.

Basic Analysis

The main function of lcplants is `LCP`. The function taxes your names(s) as input argument and will match it against the Leipzig plant Catalogue. LCP uses fuzzy matching, so that it can identify the accepted name of misspelled species as well. The `LCP` function has some options to customize the fuzzy matching and the

output, which are described below. You can get help on all arguments using `?LCP`. To run a basic analyses, all you need to do is to call `LCP` on your species name. **By default `LCP` uses direct matches only; see below if you want to use fuzzy matching**.

```
LCP("Hibiscus vitifolius")
```

```
## serial path, for searching a single taxon

## Search based on genus and epithet name

## Time difference of 3.703238 secs

## ---------------------------------------------------------

## -    End of Leipzig Catalogue of Plants (LCP) search    -

## ---------------------------------------------------------

##   ID      Submitted_Name    Order    Family    Genus    Species Infrasp
## 1  1 Hibiscus vitifolius Malvales Malvaceae Hibiscus vitifolius species
##   Infraspecies Authors Status     LCP_Accepted_Taxon PL_Comparison
## 1                   L.  valid Hibiscus vitifolius L.     identical
##   PL_Alternative   Score Insertion Deletion Substitution
## 1                matched         0        0            0
```

To assign the results to an object use the following.

```
resol <- LCP("Hibiscus vitifolius")
resol
```

The name may (and ideally should) also include the authorities.

```
LCP("Hibiscus abelmoschus var. betulifolius Mast.")
```

`LCP` also works on vectors of names. With more than 5000 names, computation may take some time.

```
LCP(c("Hibiscus abelmoschus var. betulifolius Mast.",
      "Hibiscus abutiloides Willd.",
      "Hibiscus aculeatus",
      "Hibiscus acuminatus"))
```

You can also pick a .csv file from your hard disk.

```
LCP("list")
```

Output

The output value of `LCP` is a `data.frame` with the matching information for your species names. This includes the status of the submitted name ("valid", "synonym", "invalid"), in case of synonyms the matched accepted name in LCP, a comparison to the status of the matched name in The Plant List and information on the fuzzy matching (e.g. in cases your name was misspelled, information by how match so).

Output column name	Explanation
ID	A unique identifier of the matched name
Submitted_Name	The submitted name
Order	The matched taxonomic order for the submitted name
Family	The matched taxonomic family
Genus	The matched genus
Species	The matched epitheton

Output column name	Explanation
Infrasp	The infra-specific rank, in case the matched name is an infra-specific rank
Infraspecies	The infra-specific name if applicable
Authors	The authorities of the matched name
LCP_Accepted_Taxon	The accepted name of the matched name in LCP
PL_Comparison	The status of a comparison to The Plant List, i.e. is the accepted name the same in TPL?
PL_Alternative	In case the accepted name is different in TPL, the accepted name in TPL
Score	Summary of the fuzzy matching
Insertion	In case the matching was not perfect, the number of inserted characters in the matched name
Deletion	In case the fuzzy match was not perfect, the number of deleted characters
Substitution	In case the fuzzy match was not perfect, the number of substituted characters

Further functionalities

There are several arguments to customize the matching and output of `LCP`. See `?LCP` for a description.

Change distance of the fuzzy matching

In case you expect misspellings in your names list (very likely, at least for longer lists), you might want to use fuzzy matching for the names resolution. If fuzzy matching is activated, `LCP` will also match names that differ in spelling from the submitted name. You can use the `max.distance` argument to define in how many positions the matched names are allowed to differ from the submitted name. This is especially relevant if your names include authorities, since they often differ in spelling (e.g., if spaces are used or not). One or two might be reasonable values here. Increases in `max.distance` will increase the computation time for the name resolution. In any case, it is advisable to check the results of fuzzy matching since there might be accepted names that only differ by few characters.

```r
# no fuzzy matching does not find misspelled names
fuzz <- LCP("Hibiscus vitifolios")
```

```
## Time difference of 10.3079 secs
```

```r
fuzz$Score
```

```
## [1] Epithet name not found
## Levels: Epithet name not found
```

```r
# fuzzy matching does find it
fuzz <- LCP("Hibiscus vitifolios", max.distance = 1)
```

```
## Time difference of 22.55931 secs
```

```r
fuzz$Score
```

```
## [1] misspelling: Epithet
## Levels: misspelling: Epithet
```

```r
#Also works for larger distances
fuzz <- LCP("Hibiscus vitifulios", max.distance = 2)
```

```
## Time difference of 31.2499 secs
```

```r
fuzz$Score
```

```
## [1] misspelling: Epithet
## Levels: misspelling: Epithet
```

```r
# But results become less reliable with larger distances
fuzz <- LCP("Hibiscus acetosulla", max.distance = 5)
```

```
## Time difference of 39.26238 secs
```

```r
fuzz
```

```
##   ID      Submitted_Name    Order    Family    Genus    Species Infrasp
## 1  1 Hibiscus acetosulla Malvales Malvaceae Hibiscus acetosella species
## 2  1 Hibiscus acetosulla Malvales Malvaceae Hibiscus acicularis species
##   Infraspecies        Authors Status                 LCP_Accepted_Taxon
## 1              Welw. ex Hiern  valid Hibiscus acetosella Welw. ex Hiern
## 2                     Standl.  valid        Hibiscus acicularis Standl.
##   PL_Comparison PL_Alternative                Score Insertion Deletion
## 1     identical                misspelling: Epithet         0        0
## 2     identical                misspelling: Epithet         3        3
##   Substitution
## 1            1
## 2            2
```

Run fuzzy matching on the genus

By default `LCP` will only use fuzzy matching for the species epitheton, infra-specific names and the authorities. If you want to include the genus name, use the `genus_search` argument.

```r
# no fuzzy matching does not find misspelled names
fuzz <- LCP("Hubiscus vitifolius")
```

```
## Time difference of 0.06682897 secs
```

```r
fuzz$Score
```

```
## [1] Genus name not found
## Levels: Genus name not found
```

```r
# fuzzy matching does find it
fuzz <- LCP("Hubiscus vitifolius", max.distance = 1, genus_search = TRUE)
```

```
## Time difference of 16.80288 secs
```

```r
fuzz$Score
```

```
## [1] misspelling: Genus
## Levels: misspelling: Genus
```

Parallelize computation

The name resolution can take a while to compute, especially for long lists and high matching distances (can be up to 30 minutes). TO speed up the process, you can use multiple cores if available using the `max.cores` argument. By default `LCP` will use all available cores but one.

Get all synonyms for submitted names

By default, `LCP` only returns accepted names, but you can use the `LCP` function to obtain all synonyms for your names of interest using the `status` argument.

```r
LCP("Hibiscus vitifolius", status = FALSE)
```

```
## serial path, for searching a single taxon

## Search based on genus and epithet name

## Time difference of 0.7090609 secs

## ---------------------------------------------------------

## -    End of Leipzig Catalogue of Plants (LCP) search    -

## ---------------------------------------------------------

##   ID      Submitted_Name    Order    Family    Genus    Species Infrasp
## 1  1 Hibiscus vitifolius Malvales Malvaceae Hibiscus vitifolius species
## 2  1 Hibiscus vitifolius Malvales Malvaceae Hibiscus vitifolius species
##   Infraspecies Authors  Status     LCP_Accepted_Taxon PL_Comparison
## 1                   L.   valid Hibiscus vitifolius L.     identical
## 2                Mill. synonym Hibiscus cannabinus L.     identical
##   PL_Alternative   Score Insertion Deletion Substitution
## 1                matched         0        0            0
## 2                matched         0        0            0
```

Get all names from a genus

You can also use the `LCP` function to get a list of all species names within a genus in the Leipzig Plant Catalogue.

```r
LCP("Hibiscus", genus_tab = TRUE)
```

Get all infra-specific names in a species

Similarly, you can use the `LCP` function to obtain all infra-specific namers of a species

```r
LCP("Hibiscus vitifolius", infraspecies_tab = TRUE)
```