1 Matching species tables with different taxonomic resolution

Trait-based approaches are a promising tool in ecology. Unlike taxonomy-based methods, traits may not be constrained to biogeographic boundaries [**?**] and have potential to disentangle the effects of multiple stressors [**?**].

To analyse trait-composition abundance data must be matched with trait databases like [**?**]. However these two datatables may contain species information on different taxonomic levels and perhaps data must be aggregated to a joint taxonomic level.

taxize can help in this data-cleaning step, providing a reproducible workflow. Here we illustrate this on a small fictitious example.

Suppose we have fuzzy coded trait table with 2 traits with 3 respectively 2 modalities:

```
(traits <- read.table(header = TRUE, sep = ';', stringsAsFactors=FALSE,
                      text = 'taxon;T1M1;T1M2;T1M3;T2M1;T2M2
Gammarus sp.;0;0;3;1;3
Potamopyrgus antipodarum;1;0;3;1;3
Coenagrion sp.;3;0;1;3;1
Enallagma cyathigerum;0;3;1;0;3
Erythromma sp.;0;0;3;3;1
Baetis sp.;0;0;0;0;0
'))
```

	taxon	T1M1	T1M2	T1M3	T2M1	T2M2
1	Gammarus sp.	0	0	3	1	3
2	Potamopyrgus antipodarum	1	0	3	1	3
3	Coenagrion sp.	3	0	1	3	1
4	Enallagma cyathigerum	0	3	1	0	3
5	Erythromma sp.	0	0	3	3	1
6	Baetis sp.	0	0	0	0	0

And want to match this to a table with abundances:

```
(abundances <- read.table(header = TRUE, sep = ';', stringsAsFactors=FALSE,
                          text = 'taxon;abundance;sample
Gammarus roeseli;5;1
Gammarus roeseli;6;2
Gammarus tigrinus;7;1
Gammarus tigrinus;8;2
Coenagrionidae;10;1
Coenagrionidae;6;2
Potamopyrgus antipodarum;10;1
xxxxx;10;2
'))
```

	taxon	abundance	sample
1	Gammarus roeseli	5	1
2	Gammarus roeseli	6	2
3	Gammarus tigrinus	7	1
4	Gammarus tigrinus	8	2
5	Coenagrionidae	10	1

```
6           Coenagrionidae         6      2
7 Potamopyrgus antipodarum        10      1
8                    xxxxx        10      2
```

First we do some basic data-cleaning and create a lookup-table, that will link taxa in trait table with the abundance table.

```
# first we remove ' sp.' from out trait table:
traits$taxon_cleaned <- tolower(gsub(" sp.", "", traits$taxon))

# since abundance tables can be very long with repeating taxa, we look
# only at unique taxon names This will be a lookup-table linking taxon
# names between both tables
lookup <- data.frame(taxon = tolower(unique(abundances$taxon)), stringsAsFactors = FALSE)
```

The we query the taxonomic hierarchy for both tables, this will be the backbone of this procedure:

```
library(taxize)
traits_classi <- classification(get_uid(traits$taxon_cleaned))
lookup_classi <- classification(get_uid(lookup$taxon))
```

First we look if we can find any direct matches between taxon names:

```
# first search for direct matches
direct <- match(lookup$taxon, traits$taxon_cleaned)
# and add the matched name to our lookup table
lookup$traits <- tolower(traits$taxon[direct])
lookup$match <- ifelse(!is.na(direct), "direct", NA)
lookup

                     taxon                   traits  match
1         gammarus roeseli                     <NA>   <NA>
2        gammarus tigrinus                     <NA>   <NA>
3           coenagrionidae                     <NA>   <NA>
4 potamopyrgus antipodarum potamopyrgus antipodarum direct
5                    xxxxx                     <NA>   <NA>
```

We found a direct match - *potamopyrgus antipodarum* - so nothing to do here.

Next we look for species which are on a higher taxonomic resolution than our trait table. For these species we will take directly the trait-data since no better information is available.

```
# look for cases where taxonomic resolution in abundance data is higher
# than in trait data: here we take the trait-values for the lower
# resolution
for (i in which(is.na(lookup$traits))) {
    if (is.data.frame(lookup_classi[[i]])) {
        matches <- tolower(lookup_classi[[i]]$ScientificName) %in% traits$taxon_cleaned
        if (any(matches)) {
            lookup$traits[i] <- tolower(lookup_classi[[i]]$ScientificName[matches])
            lookup$match[i] <- lookup_classi[[i]]$Rank[matches]
```

```
        }
    }
}
lookup
```

```
                      taxon                    traits  match
1         gammarus roeseli                  gammarus  genus
2        gammarus tigrinus                  gammarus  genus
3           coenagrionidae                      <NA>   <NA>
4 potamopyrgus antipodarum potamopyrgus antipodarum direct
5                    xxxxx                      <NA>   <NA>
```

So our abundance data has two *Gammarus* species, however trait data is only on genus level.

The next step is to search for species were we have to aggregate trait-data, since our abundance data is on a lower taxonomic level. We are walking the taxonomic latter for the species in our trait-data upwards and search for matches with out abundance data. Since we'll have many taxa in the trait-data belonging to one taxon, we'll take the median modality scores as an approximation. Of course also other methods may be used here, e.g. weighting by genetic distance.

```
# look for cases taxonomic resolution in abundance data is lower than in
# trait data, here we need to aggregate the trait-values (eg. median value
# for modality)

for (i in which(is.na(lookup$traits))) {
    # find matches
    agg <- sapply(traits_classi, function(x) any(tolower(x$ScientificName) %in%
        lookup$taxon[i]))
    if (sum(agg) > 1) {
        # add taxon as aggregate to trait-table
        traits <- rbind(traits, c(paste0(lookup$taxon[i], "-aggregated"), apply(traits[agg,
            2:6], 2, median), paste0(lookup$taxon[i], "-aggregated")))
        # fill lookup table
        lookup$traits[i] <- paste0(lookup$taxon[i], "-aggregated")
        lookup$match[i] <- "aggregated"
    }
}
lookup
```

```
##                       taxon                      traits      match
## 1         gammarus roeseli                    gammarus      genus
## 2        gammarus tigrinus                    gammarus      genus
## 3           coenagrionidae coenagrionidae-aggregated aggregated
## 4 potamopyrgus antipodarum  potamopyrgus antipodarum     direct
## 5                    xxxxx                        <NA>       <NA>
```

Finally we have only one taxon left - clearly an error. We remove this from our dataset:

```
abundances <- abundances[!abundances$taxon == lookup$taxon[is.na(lookup$traits)],
    ]
```

No we can create *species x sites* and *traits x species* matrices, which could be plugged into methods to analyse trait responses [**?**].

```
# species (as matched with trait table) by site matrix
abundances$traits_taxa <- lookup$traits[match(tolower(abundances$taxon), lookup$taxon)]

library(reshape2)
# reshape data to long format and name rows by samples
L <- dcast(abundances, sample ~ traits_taxa, fun.aggregate = sum, value.var = "abundance")
rownames(L) <- L$sample
L$sample <- NULL
L

##   coenagrionidae-aggregated gammarus potamopyrgus antipodarum
## 1                        10       12                       10
## 2                         6       14                        0


# traits by species matrix
Q <- traits[, 2:7][match(names(L), traits$taxon_cleaned), ]
rownames(Q) <- Q$taxon_cleaned
Q$taxon_cleaned <- NULL
Q

##                           T1M1 T1M2 T1M3 T2M1 T2M2
## coenagrionidae-aggregated    0    0    1    3    1
## gammarus                     0    0    3    1    3
## potamopyrgus antipodarum     1    0    3    1    3


# check
all(rownames(Q) == colnames(L))

## [1] TRUE
```

This is just an example how taxonomic APIs (via taxize) could be used to search for matches up- and downwards the taxonomic ladder. We are looking forward to integrate the freshwaterecology.info database [**?**] into taxize, which will facilitate trait-based analyses in R.