

taxa: An R package for taxonomic data

Zachary Foster, Scott Chamberlain, and Niklaus Grunwald

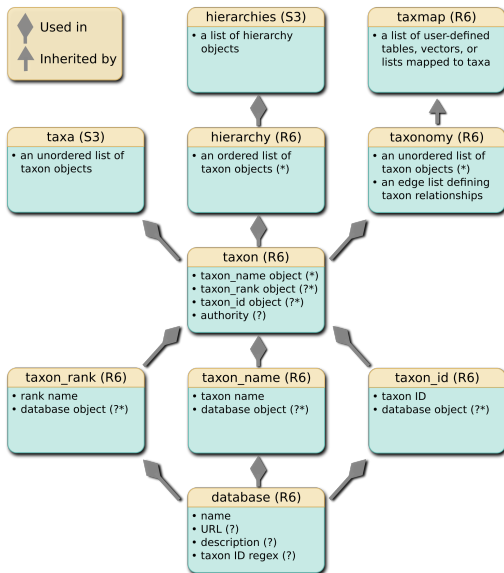
The challenges of taxonomic data

- ▶ Taxonomic data is hierarchical
- ▶ It is often associated with other data
- ▶ "Taxa" can be names, classifications of names, or IDs
- ▶ Each source of taxonomic data formats things differently

Database	FASTA sequence header format
UNITE	Lachnum_sp JQ347180 SH189775.06FU reps k__Fungi;p__Ascomycota;c__Leotiomycetes;o__Helotial...
Genbank	gil626414534 ref NR_119473.1 Lysurus cruciatus MA Fungi 26792 ITS region; from TYPE mater...
PR2	10-044.1.1773 Eukaryota Stramenopiles Stramenopiles_X Oomycota Oomycota_X Oomycota_XX Oomy...
RDP	S001191995 uncultured archaeon; LCDARCH35 Lineage=Root;rootrank;Archaea;domain;"Euryarchaeo...
ITS1	AF455489_ITS1_GB Lecanicillium aphanocladii 132584 ITS1 located by Genbank annotation, 18...

What taxa provides

- ▶ Classes to hold taxa, taxonomies, and associated data
- ▶ Flexible parsers to convert raw data to these classes
- ▶ Dplyr-inspired functions to manipulate these classes
- ▶ A flexible base for other packages to use



taxmap: user-defined data mapped to a taxonomy

```
> ex_taxmap
<Taxmap>
17 taxa: b. Mammalia, c. Plantae, d. Felidae ... q. lycopersicum, r. tuberosum
17 edges: NA->b, NA->c, b->d, b->e, b->f, c->g ... i->n, j->o, k->p, l->q, l->r
4 data sets:
  info:
    # A tibble: 6 x 4
      taxon_id name  n_legs dangerous
    <chr>    <chr>  <dbl> <lgl>
1 m      tiger    4. TRUE
2 n      cat      4. FALSE
3 o      mole     4. FALSE
    # ... with 3 more rows
  phylopic_ids: a named vector of 'character' with 6 items
    m. e148eabb-f138-43[truncated] ... r. 63604565-0406-46[truncated]
  foods: a list of 6 items named by taxa:
    m, n, o, p, q, r
  abund:
    # A tibble: 8 x 5
      taxon_id code  sample_id count taxon_index
    <chr>    <fct> <fct>      <dbl>      <int>
1 m      T    A          1.          1
2 n      C    A          2.          2
3 o      M    B          5.          3
    # ... with 5 more rows
1 functions:
  reaction
```

Reading data from diverse formats

Input type

Input data format

	Simple	Embedded	Raw string
	<pre>> print(data) [1] "input_1" "input_2" [3] "input_3"</pre>	<pre>> print(data) x input y 1 a input_1 100 2 b input_2 200 3 c input_3 300</pre>	<pre>> print(data) [1] ">id:a-tax:input_1" [2] ">id:b-tax:input_2" [3] ">id:c-tax:input_3"</pre>
Classification Primates;Hominidae;Homo;sapiens	<pre>> print(data) [1] "Primates;Hominidae;Hom..." [2] "Primates;Haplorhini;Cr..." > parse_tax_data(data, class_sep = ";")</pre>	<pre>> print(data) x class y 1 a Primates;Hominidae;... 100 2 b Primates;Haplorhini... 200 > parse_tax_data(data, class_cols = "class", class_sep = ";")</pre>	<pre>> print(data) [1] ">id:a-tax:Primates;Hom..." [2] ">id:b-tax:Primates;Hap..." > extract_tax_data(data, regex = ">id:(.+)-tax:(.+)", key = c("info", "class"), class_sep = ";")</pre>
Taxon ID 9606	<pre>> print(data) [1] "9606" "100937" ... > lookup_tax_data(data, type = "taxon_id")</pre>	<pre>> print(data) x id y 1 a 9606 100 2 b 100937 200 > lookup_tax_data(data, type = "taxon_id", column = "id")</pre>	<pre>> print(data) [1] ">id:a-tax:9606" [2] ">id:b-tax:100937" > extract_tax_data(data, regex = ">id:(.+)-tax:(.+)", key = c("info", "taxon_id"), database = "ncbi")</pre>
Taxon name Homo sapiens	<pre>> print(data) [1] "Homo sapiens" [2] "Primates" ... > lookup_tax_data(data, type = "taxon_name")</pre>	<pre>> print(data) x name y 1 a Homo sapiens 100 2 b Primates 200 > lookup_tax_data(data, type = "taxon_name", column = "name")</pre>	<pre>> print(data) [1] ">id:a-tax:Homo sapiens" [2] ">id:b-tax:Primates" > extract_tax_data(data, regex = ">id:(.+)-tax:(.+)", key = c("info", "taxon_name"), database = "ncbi")</pre>
Sequence ID AC073210	<pre>> print(data) [1] "AC073210" "KC312885" ... > lookup_tax_data(data, type = "seq_id")</pre>	<pre>> print(data) x ncbi_id y 1 a AC073210 100 2 b KC312885 200 > lookup_tax_data(data, type = "seq_id", column = "ncbi_id")</pre>	<pre>> print(data) [1] ">id:a-tax:AC073210" [2] ">id:b-tax:KC312885" > extract_tax_data(data, regex = ">id:(.+)-tax:(.+)", key = c("info", "seq_id"), database = "ncbi")</pre>

Dplyr-like manipulation of taxonomic data

Subset taxonomy and data to one taxon:

```
filter_taxa(x, taxon_names == "Plantae", subtaxa = TRUE)
```

Subset taxonomy to one rank:

```
filter_taxa(x, taxon_ranks == "genus", supertaxa = TRUE)
```

Subset data and remove any taxa not in subset:

```
filter_obs(x, "info", n_legs == 4, drop_taxa = TRUE)
```

Add a column to a dataset:

```
mutate_obs(x, "info", bipedal = n_legs == 2)
```