Phylogeny Statistics

Part I – Phylogeny Content Statistics

Packages

```
library(dplyr)
library(pbapply)
```

Data

The data used for the descriptive statistics of the phylogeny comes from two sources: the overall dataset used to build the phylogeny and a tab-delimited table exported by Sequence Matrix v1.8 containing the names of the taxa used in the final alignment along with the identity and length of the gene regions sequenced for those taxa. The two data frames are joined together using `dplyr::semi_join()` to create a single data frame containing information on the family, phytogeographic domain, lifeform and location (state) for the species found in the phylogeny. The composite taxa from outside of Brazil were removed for these statistics.

```
# data frame containing information for all species found in Brazil

data <- read.csv(file.path('.', 'nordeste.csv'), header = TRUE)

# data frame containing all taxa found in the phylogeny

phylo <- data.frame(read.csv(file.path('.', 'alignment_table.csv')))
phylo <- data.frame(grep('OUTSIDE', phylo$Taxon, value = TRUE, invert = TRUE))
  # removes OUTSIDE taxa from the data frame; total of 366 OUTSIDE taxa removed
names(phylo) <- 'Taxon'
phylo$Combination <- sub('.*aceae ', '', phylo$Taxon)

stats <- semi_join(data, phylo, by = 'Combination')
unknowns <- data.frame(base::setdiff(phylo$Combination, stats$Combination))
  # 6 taxa were renamed during the ncbi search and their synonym equivalent
  # in the data set could not be found; they are tacked on to the end of
  # the following stats data set for completeness
names(unknowns) <- 'Combination'
unknowns
unknowns$Family <- c('Bromeliaceae', 'Fabaceae', 'Fabaceae', 'Fabaceae', 'Poaceae', 'Poaceae')
stats <- full_join(stats, unknowns)
```

```
## Joining, by = c("Family", "Combination")
```

The `tblFunc` Function

The function `tblFunc` was created to count instances of a desired variable in both the overall set of data and the subset containing only the species found in the phylogeny. `tblFunc` also provides the percentage value of the variable to the whole in the data.Percentage and tree.Percentage columns. Lastly, `tblFunc` creates a column (total.Percentage) indicating the percentage of overall data that is found in the phylogeny for a particular variable.

```r
tblFunc <- function(x, y, States = FALSE){ # function to create % tables

    # x -- variable from the overall set of data
    # y -- variable from the phylogeny subset
    # States -- whether or not the variable of interest is the location (state)

    # data frame containing counts and percentages of overall data

    tbl.x <- x
    res.x <- data.frame(cbind(row.names(tbl.x), tbl.x,
      if(States == TRUE) {round(tbl.x / length(data[, 1]) * 100, 2)
      } else {
        round(prop.table(tbl.x) * 100, 2)}))
    colnames(res.x) <- c('id', 'data.Count', 'data.Percentage')

    # data frame containing counts and percentages of phylogeny data

    tbl.y <- y
    res.y <- data.frame(cbind(row.names(tbl.y), tbl.y,
      if(States == TRUE) {
      round(tbl.y / length(stats[, 1]) * 100, 2)
        } else {
          round(prop.table(tbl.y) * 100, 2)}))
    colnames(res.y) <- c('id', 'tree.Count', 'tree.Percentage')

    # join the above data frames and add a column showing the
    # percentage of overall data that is found in the tree

    res <- full_join(res.x, res.y, by = 'id') # combine the two based on the 'id' column
    res$id <- sub('^$', 'Unknown', res$id) # renames any empty id names 'Unknown'
    res[is.na(res)] <- 0
    res <- mutate_all(res, type.convert) # converts columns to the appropriate type
    res$total.Percentage <- round(res$tree.Count / res$data.Count * 100, 2)
      # column showing the percentage of overall data that is found in the tree
    res
}
```

Descriptive Statistics

Family

We can use the Family columns of our data sets to illustrate the use of `tblFunc`. The two variables for the function are `table(data$Family)` and `table(stats$Family)` and these provide a count of the number of taxa belonging to each family in the data sets and the function does the rest. As the table shows, in the overall data, there are 1,328 members of the Fabaceae family, comprising 8.12% of all taxa in the data set. There are 601 members of the Acanthaceae family in the phylogeny, comprising 14.17% of all taxa in the tree. Of all 1,328 Acanthaceae in the overall data set, only 45.26% of them were used in the creation of the phylogeny. The function has also been applied to the Lifeform, Phyotogeographic Domain and State columns below. Please note that the taxa from outside of Brazil were not included in these statistics.

Table 1: Family – A table showing the count and percentages of taxa per family. data.Count: Number of taxa in the overall data from family X. data.Percentage: Percentage of overall data composed of taxa from family X. tree.Count: Number of taxa in the tree family X. tree.Percentage: Percentage of the tree composed of taxa from family X. total.Percentage: Percentage of all taxa from family X used in the tree.

	id	data.Count	data.Percentage	tree.Count	tree.Percentage	total.Percentage
77	Fabaceae	1328	8.12	601	14.17	45.26
133	Orchidaceae	1174	7.18	388	9.15	33.05
146	Poaceae	722	4.42	299	7.05	41.41
20	Asteraceae	1272	7.78	289	6.81	22.72
31	Bromeliaceae	635	3.88	190	4.48	29.92
12	Apocynaceae	405	2.48	150	3.54	37.04
110	Malpighiaceae	306	1.87	124	2.92	40.52
26	Bignoniaceae	268	1.64	117	2.76	43.66
116	Melastomataceae	708	4.33	107	2.52	15.11
64	Cyperaceae	372	2.28	104	2.45	27.96
125	Myrtaceae	524	3.21	98	2.31	18.70
163	Rubiaceae	552	3.38	82	1.93	14.86
79	Gesneriaceae	72	0.44	64	1.51	88.89
76	Euphorbiaceae	496	3.03	63	1.49	12.70
35	Cactaceae	152	0.93	60	1.41	39.47
111	Malvaceae	399	2.44	60	1.41	15.04
177	Solanaceae	243	1.49	57	1.34	23.46
98	Lamiaceae	279	1.71	52	1.23	18.64
136	Passifloraceae	80	0.49	49	1.16	61.25
15	Araceae	168	1.03	46	1.08	27.38
10	Annonaceae	115	0.70	44	1.04	38.26
143	Piperaceae	206	1.26	44	1.04	21.36
99	Lauraceae	180	1.10	37	0.87	20.56
102	Lentibulariaceae	61	0.37	37	0.87	60.66
58	Convolvulaceae	266	1.63	36	0.85	13.53
67	Dioscoreaceae	75	0.46	36	0.85	48.00
170	Sapotaceae	95	0.58	35	0.83	36.84
48	Celastraceae	78	0.48	33	0.78	42.31
17	Arecaceae	118	0.72	30	0.71	25.42
122	Moraceae	92	0.56	26	0.61	28.26
148	Polygalaceae	117	0.72	25	0.59	21.37
169	Sapindaceae	205	1.25	25	0.59	12.20
139	Phyllanthaceae	75	0.46	24	0.57	32.00
66	Dilleniaceae	36	0.22	23	0.54	63.89
6	Alstroemeriaceae	28	0.17	22	0.52	78.57
51	Chrysobalanaceae	82	0.50	21	0.50	25.61
101	Lecythidaceae	29	0.18	21	0.50	72.41
195	Violaceae	41	0.25	21	0.50	51.22
94	Iridaceae	80	0.49	19	0.45	23.75
194	Verbenaceae	164	1.00	18	0.42	10.98
18	Aristolochiaceae	48	0.29	17	0.40	35.42
149	Polygonaceae	51	0.31	17	0.40	33.33
167	Salicaceae	51	0.31	17	0.40	33.33

Table 1: Family (continued)

	id	data.Count	data.Percentage	tree.Count	tree.Percentage	total.Percentage
112	Marantaceae	100	0.61	16	0.38	16.00
5	Alismataceae	25	0.15	15	0.35	60.00
60	Cucurbitaceae	82	0.50	15	0.35	18.29
73	Eriocaulaceae	507	3.10	15	0.35	2.96
192	Urticaceae	38	0.23	15	0.35	39.47
7	Amaranthaceae	100	0.61	14	0.33	14.00
8	Amaryllidaceae	47	0.29	14	0.33	29.79
29	Boraginaceae	91	0.56	14	0.33	15.38
150	Pontederiaceae	18	0.11	14	0.33	77.78
181	Symplocaceae	25	0.15	14	0.33	56.00
117	Meliaceae	35	0.21	13	0.31	37.14
147	Podostemaceae	22	0.13	13	0.31	59.09
9	Anacardiaceae	29	0.18	12	0.28	41.38
56	Commelinaceae	65	0.40	12	0.28	18.46
144	Plantaginaceae	74	0.45	12	0.28	16.22
38	Campanulaceae	31	0.19	10	0.24	32.26
118	Menispermaceae	31	0.19	10	0.24	32.26
128	Ochnaceae	83	0.51	10	0.24	12.05
131	Onagraceae	33	0.20	10	0.24	30.30
174	Simaroubaceae	18	0.11	10	0.24	55.56
1	Acanthaceae	188	1.15	9	0.21	4.79
14	Aquifoliaceae	36	0.22	9	0.21	25.00
33	Burseraceae	27	0.17	9	0.21	33.33
55	Combretaceae	31	0.19	9	0.21	29.03
68	Droseraceae	23	0.14	9	0.21	39.13
151	Portulacaceae	13	0.08	9	0.21	69.23
16	Araliaceae	34	0.21	8	0.19	23.53
52	Cleomaceae	21	0.13	8	0.19	38.10
89	Hydrocharitaceae	12	0.07	8	0.19	66.67
126	Nyctaginaceae	33	0.20	8	0.19	24.24
140	Phytolaccaceae	19	0.12	8	0.19	42.11
54	Clusiaceae	33	0.20	7	0.17	21.21
105	Loasaceae	8	0.05	7	0.17	87.50
108	Lythraceae	120	0.73	7	0.17	5.83
153	Primulaceae	67	0.41	7	0.17	10.45
156	Quiinaceae	8	0.05	7	0.17	87.50
165	Rutaceae	102	0.62	7	0.17	6.86
193	Velloziaceae	193	1.18	7	0.17	3.63
198	Vochysiaceae	49	0.30	7	0.17	14.29
45	Caricaceae	6	0.04	6	0.14	100.00
72	Ericaceae	56	0.34	6	0.14	10.71
106	Loganiaceae	71	0.43	6	0.14	8.45
129	Olacaceae	19	0.12	6	0.14	31.58
11	Apiaceae	31	0.19	5	0.12	16.13
36	Calophyllaceae	32	0.20	5	0.12	15.62
69	Ebenaceae	17	0.10	5	0.12	29.41

Table 1: Family (continued)

	id	data.Count	data.Percentage	tree.Count	tree.Percentage	total.Percentage
160	Rhamnaceae	27	0.17	5	0.12	18.52
24	Begoniaceae	61	0.37	4	0.09	6.56
42	Capparaceae	14	0.09	4	0.09	28.57
91	Hypericaceae	16	0.10	4	0.09	25.00
104	Linderniaceae	6	0.04	4	0.09	66.67
124	Myristicaceae	12	0.07	4	0.09	33.33
152	Potamogetonaceae	9	0.06	4	0.09	44.44
189	Turneraceae	98	0.60	4	0.09	4.08
32	Burmanniaceae	11	0.07	3	0.07	27.27
34	Cabombaceae	4	0.02	3	0.07	75.00
41	Cannaceae	3	0.02	3	0.07	100.00
70	Elaeocarpaceae	12	0.07	3	0.07	25.00
74	Erythroxylaceae	88	0.54	3	0.07	3.41
75	Escalloniaceae	6	0.04	3	0.07	50.00
86	Hernandiaceae	6	0.04	3	0.07	50.00
93	Icacinaceae	9	0.06	3	0.07	33.33
97	Lacistemataceae	6	0.04	3	0.07	50.00
114	Martyniaceae	3	0.02	3	0.07	100.00
120	Molluginaceae	4	0.02	3	0.07	75.00
121	Monimiaceae	18	0.11	3	0.07	16.67
138	Peraceae	9	0.06	3	0.07	33.33
154	Proteaceae	18	0.11	3	0.07	16.67
161	Rhizophoraceae	4	0.02	3	0.07	75.00
162	Rosaceae	10	0.06	3	0.07	30.00
175	Siparunaceae	7	0.04	3	0.07	42.86
190	Typhaceae	3	0.02	3	0.07	100.00
200	Xyridaceae	117	0.72	3	0.07	2.56
19	Asparagaceae	8	0.05	2	0.05	25.00
27	Bixaceae	5	0.03	2	0.05	40.00
30	Brassicaceae	2	0.01	2	0.05	100.00
40	Cannabaceae	5	0.03	2	0.05	40.00
44	Cardiopteridaceae	4	0.02	2	0.05	50.00
61	Cunoniaceae	7	0.04	2	0.05	28.57
83	Haemodoraceae	2	0.01	2	0.05	100.00
95	Juncaceae	4	0.02	2	0.05	50.00
107	Loranthaceae	62	0.38	2	0.05	3.23
113	Marcgraviaceae	11	0.07	2	0.05	18.18
119	Menyanthaceae	2	0.01	2	0.05	100.00
127	Nymphaeaceae	9	0.06	2	0.05	22.22
130	Oleaceae	9	0.06	2	0.05	22.22
134	Orobanchaceae	29	0.18	2	0.05	6.90
135	Oxalidaceae	51	0.31	2	0.05	3.92
157	Ranunculaceae	9	0.06	2	0.05	22.22
168	Santalaceae	42	0.26	2	0.05	4.76
186	Trigoniaceae	12	0.07	2	0.05	16.67
196	Vitaceae	29	0.18	2	0.05	6.90

Table 1: Family (continued)

	id	data.Count	data.Percentage	tree.Count	tree.Percentage	total.Percentage
2	Achariaceae	8	0.05	1	0.02	12.50
3	Adoxaceae	1	0.01	1	0.02	100.00
4	Aizoaceae	1	0.01	1	0.02	100.00
22	Basellaceae	2	0.01	1	0.02	50.00
23	Bataceae	1	0.01	1	0.02	100.00
28	Bonnetiaceae	1	0.01	1	0.02	100.00
46	Caryocaraceae	6	0.04	1	0.02	16.67
49	Ceratophyllaceae	1	0.01	1	0.02	100.00
50	Chloranthaceae	1	0.01	1	0.02	100.00
53	Clethraceae	1	0.01	1	0.02	100.00
57	Connaraceae	31	0.19	1	0.02	3.23
59	Costaceae	6	0.04	1	0.02	16.67
62	Cyclanthaceae	7	0.04	1	0.02	14.29
63	Cymodoceaceae	2	0.01	1	0.02	50.00
65	Dichapetalaceae	6	0.04	1	0.02	16.67
80	Goodeniaceae	1	0.01	1	0.02	100.00
81	Goupiaceae	1	0.01	1	0.02	100.00
84	Haloragaceae	2	0.01	1	0.02	50.00
85	Heliconiaceae	12	0.07	1	0.02	8.33
92	Hypoxidaceae	3	0.02	1	0.02	33.33
103	Linaceae	4	0.02	1	0.02	25.00
109	Magnoliaceae	1	0.01	1	0.02	100.00
115	Mayacaceae	4	0.02	1	0.02	25.00
123	Myoporaceae	1	0.01	1	0.02	100.00
137	Pentaphylacaceae	5	0.03	1	0.02	20.00
141	Picramniaceae	14	0.09	1	0.02	7.14
145	Plumbaginaceae	1	0.01	1	0.02	100.00
158	Rapateaceae	4	0.02	1	0.02	25.00
164	Ruppiaceae	1	0.01	1	0.02	100.00
171	Schlegeliaceae	1	0.01	1	0.02	100.00
178	Stemonuraceae	1	0.01	1	0.02	100.00
179	Styracaceae	15	0.09	1	0.02	6.67
180	Surianaceae	1	0.01	1	0.02	100.00
183	Theaceae	1	0.01	1	0.02	100.00
185	Thymelaeaceae	8	0.05	1	0.02	12.50
191	Ulmaceae	3	0.02	1	0.02	33.33
201	Zingiberaceae	5	0.03	1	0.02	20.00
202	Zygophyllaceae	2	0.01	1	0.02	50.00
13	Apodanthaceae	2	0.01	NA	NA	NA
21	Balanophoraceae	8	0.05	NA	NA	NA
25	Berberidaceae	3	0.02	NA	NA	NA
37	Calyceraceae	1	0.01	NA	NA	NA
39	Canellaceae	1	0.01	NA	NA	NA
43	Caprifoliaceae	6	0.04	NA	NA	NA
47	Caryophyllaceae	3	0.02	NA	NA	NA
71	Elatinaceae	1	0.01	NA	NA	NA

Table 1: Family (continued)

	id	data.Count	data.Percentage	tree.Count	tree.Percentage	total.Percentage
78	Gentianaceae	57	0.35	NA	NA	NA
82	Griseliniaceae	1	0.01	NA	NA	NA
87	Humiriaceae	4	0.02	NA	NA	NA
88	Hydnoraceae	1	0.01	NA	NA	NA
90	Hydroleaceae	1	0.01	NA	NA	NA
96	Krameriaceae	4	0.02	NA	NA	NA
100	Laxmanniaceae	1	0.01	NA	NA	NA
132	Opiliaceae	1	0.01	NA	NA	NA
142	Picrodendraceae	2	0.01	NA	NA	NA
155	Putranjivaceae	2	0.01	NA	NA	NA
159	Rhabdodendraceae	1	0.01	NA	NA	NA
166	Sabiaceae	3	0.02	NA	NA	NA
172	Schoepfiaceae	2	0.01	NA	NA	NA
173	Scrophulariaceae	8	0.05	NA	NA	NA
176	Smilacaceae	27	0.17	NA	NA	NA
182	Taccaceae	1	0.01	NA	NA	NA
184	Thismiaceae	1	0.01	NA	NA	NA
187	Triuridaceae	4	0.02	NA	NA	NA
188	Tropaeolaceae	1	0.01	NA	NA	NA
197	Vivianiaceae	1	0.01	NA	NA	NA
199	Winteraceae	1	0.01	NA	NA	NA

Lifeform

Table 2: Lifeform – A table showing the count and percentages of taxa by lifeform. data.Count: N
of taxa in the overall data exhibiting lifeform X. data.Percentage: Percentage of overall data comp
taxa exhibiting lifeform X. tree.Count: Number of taxa in the tree exhibiting lifeform X. tree.Per
Percentage of the tree composed of taxa exhibiting lifeform X. total.Percentage: Percentage of
exhibiting lifeform X used in the tree.

	id	data.Count	data.Percentage	tree.C
26	Erva	5682	34.75	
18	Arvore	2155	13.18	
32	Liana/voluvel/trepadeira	1644	10.06	
3	Arbusto; Arvore	1174	7.18	
2	Arbusto	2096	12.82	
36	Subarbusto	1443	8.83	
15	Arbusto; Subarbusto	823	5.03	
29	Erva; Subarbusto	535	3.27	
12	Arbusto; Liana/voluvel/trepadeira	171	1.05	
11	Arbusto; Erva; Subarbusto	110	0.67	
1	Unknown	59	0.36	
6	Arbusto; Arvore; Subarbusto	42	0.26	
27	Erva; Liana/voluvel/trepadeira	54	0.33	
33	Liana/voluvel/trepadeira; Subarbusto	56	0.34	
4	Arbusto; Arvore; Liana/voluvel/trepadeira	40	0.24	

Table 2: Lifeform (continued)

	id	data.Count	data.Percentage	tree.C
31	Erva; Suculenta	18	0.11	
30	Erva; Subarbusto; Suculenta	12	0.07	
9	Arbusto; Erva	30	0.18	
17	Arbusto; Suculenta	24	0.15	
19	Arvore; Liana/voluvel/trepadeira	16	0.10	
37	Subarbusto; Suculenta	14	0.09	
16	Arbusto; Subarbusto; Suculenta	13	0.08	
22	Bambu	26	0.16	
7	Arbusto; Arvore; Suculenta	6	0.04	
20	Arvore; Subarbusto	10	0.06	
28	Erva; Liana/voluvel/trepadeira; Subarbusto	4	0.02	
10	Arbusto; Erva; Liana/voluvel/trepadeira; Subarbusto	4	0.02	
13	Arbusto; Liana/voluvel/trepadeira; Subarbusto	14	0.09	
14	Arbusto; Liana/voluvel/trepadeira; Suculenta	6	0.04	
24	Desconhecida	28	0.17	
35	Palmeira	31	0.19	
5	Arbusto; Arvore; Liana/voluvel/trepadeira; Subarbusto	1	0.01	
8	Arbusto; Desconhecida	1	0.01	
23	Bambu; Liana/voluvel/trepadeira	2	0.01	
25	Dracenuide; Erva	1	0.01	
34	Liana/voluvel/trepadeira; Suculenta	1	0.01	
21	Arvore; Suculenta	3	0.02	

Phytogeographic Domain

Table 3: Phytogeographic Domain – A table showing the count and percentages of taxa by phy
domain. data.Count: Number of taxa in the overall data found in domain X. data.Percentage
of overall data composed of taxa which can be found in domain X. tree.Count: Number o
tree found in domain X. tree.Percentage: Percentage of the tree composed of taxa which ca
domain X. total.Percentage: Percentage of all taxa which can be found in domain X used

	id	data.Count	data.Percentag
53	Mata Atlantica	4088	25.0
46	Cerrado, Mata Atlantica	1649	10.0
45	Cerrado	3309	20.2
5	Amazunia, Caatinga, Cerrado, Mata Atlantica	679	4.1
18	Amazunia, Cerrado, Mata Atlantica	567	3.4
2	Amazunia	591	3.6
25	Amazunia, Mata Atlantica	479	2.9
31	Caatinga, Cerrado, Mata Atlantica	570	3.4
30	Caatinga, Cerrado	869	5.3
17	Amazunia, Cerrado	481	2.9
29	Caatinga	995	6.0
7	Amazunia, Caatinga, Cerrado, Mata Atlantica, Pampa, Pantanal	181	1.1
8	Amazunia, Caatinga, Cerrado, Mata Atlantica, Pantanal	189	1.1
38	Caatinga, Mata Atlantica	329	2.0

Table 3: Phytogeographic Domain

	id	data.Count	data.Percentag
4	Amazunia, Caatinga, Cerrado	232	1.4
12	Amazunia, Caatinga, Mata Atlantica	95	0.5
1	Unknown	115	0.7
47	Cerrado, Mata Atlantica, Pampa	147	0.9
54	Mata Atlantica, Pampa	99	0.6
21	Amazunia, Cerrado, Mata Atlantica, Pantanal	72	0.4
6	Amazunia, Caatinga, Cerrado, Mata Atlantica, Pampa	46	0.2
3	Amazunia, Caatinga	69	0.4
32	Caatinga, Cerrado, Mata Atlantica, Pampa	44	0.2
24	Amazunia, Cerrado, Pantanal	37	0.2
49	Cerrado, Mata Atlantica, Pantanal	56	0.3
34	Caatinga, Cerrado, Mata Atlantica, Pantanal	39	0.2
11	Amazunia, Caatinga, Cerrado, Pantanal	30	0.1
33	Caatinga, Cerrado, Mata Atlantica, Pampa, Pantanal	22	0.1
48	Cerrado, Mata Atlantica, Pampa, Pantanal	33	0.2
15	Amazunia, Caatinga, Mata Atlantica, Pantanal	13	0.0
20	Amazunia, Cerrado, Mata Atlantica, Pampa, Pantanal	17	0.1
27	Amazunia, Mata Atlantica, Pantanal	20	0.1
19	Amazunia, Cerrado, Mata Atlantica, Pampa	16	0.1
52	Cerrado, Pantanal	40	0.2
44	Caatinga, Pantanal	12	0.0
50	Cerrado, Pampa	25	0.1
56	Mata Atlantica, Pantanal	14	0.0
39	Caatinga, Mata Atlantica, Pampa	8	0.0
26	Amazunia, Mata Atlantica, Pampa	5	0.0
28	Amazunia, Pantanal	9	0.0
37	Caatinga, Cerrado, Pantanal	20	0.1
41	Caatinga, Mata Atlantica, Pantanal	8	0.0
13	Amazunia, Caatinga, Mata Atlantica, Pampa	3	0.0
14	Amazunia, Caatinga, Mata Atlantica, Pampa, Pantanal	2	0.0
55	Mata Atlantica, Pampa, Pantanal	5	0.0
10	Amazunia, Caatinga, Cerrado, Pampa, Pantanal	1	0.0
16	Amazunia, Caatinga, Pantanal	4	0.0
35	Caatinga, Cerrado, Pampa	4	0.0
40	Caatinga, Mata Atlantica, Pampa, Pantanal	1	0.0
42	Caatinga, Pampa	1	0.0
43	Caatinga, Pampa, Pantanal	2	0.0
9	Amazunia, Caatinga, Cerrado, Pampa	1	0.0
22	Amazunia, Cerrado, Pampa	1	0.0
23	Amazunia, Cerrado, Pampa, Pantanal	1	0.0
36	Caatinga, Cerrado, Pampa, Pantanal	1	0.0
51	Cerrado, Pampa, Pantanal	1	0.0
57	Pampa	1	0.0
58	Pantanal	1	0.0

States

The State columns were arranged differently from the other variables of interest in the data set and so `tblFunc` acts slightly different for the states. Rather than using `table(x)` and `table(y)`, data frames containing column sums were created and used as the data sources for the function. Each column represented a single state and the elements inside were either a 1 (for the presence of the taxa in that state), a 0 (for the absence) or an NA (for no data). When summed together, this provided a count of how many taxa were found in each state. As taxa could be found in multiple states, the proportion of a particular state's contribution to the whole was determined by dividing the sum of that state by the total number of taxa (16,349 for the overall data and 4,489 for the phylogeny) in each data set.

Table 4: States – A table showing the count and percentages of taxa per Brazillian state. data.Count: Number of taxa in the overall data found in state X. data.Percentage: Percentage of overall data composed of taxa which can be found in state X. tree.Count: Number of taxa in the tree found in state X. tree.Percentage: Percentage of the tree composed of taxa which can be found in state X. total.Percentage: Percentage of all taxa which can be found in state X used in the tree.

	id	data.Count	data.Percentage	tree.Count	tree.Percentage	total.Percentage
13	MG	11136	68.11	2933	69.16	26.34
5	BA	8887	54.36	2780	65.55	31.28
25	SP	5741	35.12	2123	50.06	36.98
19	RJ	5211	31.87	1926	45.41	36.96
8	ES	4255	26.03	1669	39.35	39.22
16	PR	4010	24.53	1626	38.34	40.55
9	GO	4299	26.30	1558	36.74	36.24
11	MT	3646	22.30	1557	36.71	42.70
14	PA	3084	18.86	1473	34.73	47.76
17	PE	2982	18.24	1355	31.95	45.44
4	AM	2535	15.51	1314	30.98	51.83
10	MA	2805	17.16	1282	30.23	45.70
24	SC	2932	17.93	1265	29.83	43.14
12	MS	2716	16.61	1235	29.12	45.47
7	DF	2795	17.10	1152	27.16	41.22
6	CE	2360	14.44	1074	25.32	45.51
21	RS	2227	13.62	999	23.56	44.86
1	AC	1521	9.30	900	21.22	59.17
2	AL	1808	11.06	892	21.03	49.34
22	RO	1639	10.03	892	21.03	54.42
15	PB	1826	11.17	889	20.96	48.69
27	TO	1987	12.15	856	20.18	43.08
23	RR	1541	9.43	842	19.85	54.64
18	PI	1967	12.03	839	19.78	42.65
3	AP	1413	8.64	785	18.51	55.56
26	SE	1610	9.85	782	18.44	48.57
20	RN	1237	7.57	630	14.85	50.93

Part II – Missing Data and Node Support Values

Packages

```
library(dplyr)
library(phylotools)
```

Data

Determining the amount of missing data in the alignment and the average node support values required the final, combined alignment and the phylogenetic tree data, respectively.

```
# alignment data

alignment <- read.phylip(file.path
        ('..', 'results', 'alignments', 'combined_alignment', 'final_combined_alignment'))
alignment$family <- factor(gsub('_.*', '', alignment$seq.name))

# phylogenetic tree data

tree <- read.tree(file.path
        ('..', 'results', 'trees', 'combined_tree', 'combined_tree.tree'))
tree$node.label <- as.numeric(tree$node.label)
```

Missing Data

For our purposes, missing data was considered to be anything other than an A, T, C or G. This includes indels (both interior and exterior), abbreviations for uncertain nucleotides (N, R, Y, W, S, M, K, etc.) and questions marks (indiciating a complete lack of a gene region sequence). The amount of missing data in the alignment was overall rather high with a mean of 90.16% of the alignment containing missing data. Missing data could potentially be reduced by manually pruning large gaps in the alignment caused by small numbers of taxa. While the average family contained a mean of 89.26% missing data, the alignments of some families proved to be more complete than others with a range of 31.86% separating the minimum and maximum missing data percentages.

```
# create a column displaying the number of missing values per species

alignment$missing <- pbsapply(alignment$seq.text, function(x) {
  length(unlist(regmatches(x, gregexpr('[^ATCG]', x))))
}) # 20,376 total loci per taxa

summary(alignment$missing) / 20376 * 100 # summary of missing percentages
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   63.54   88.09   91.43   90.16   93.72   97.55

ali_mis <- alignment %>%
    group_by(family) %>%
    summarise(missing.perc = ((sum(missing) / length(family)) / 20376) * 100)
summary(ali_mis$missing.perc) # summary of missing percentages by family
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   65.43   85.86   89.26   88.19   92.42   97.29
```

Node Support Values

Node support values had a median of 70.0 and a mean of 63.8. A fairly large number of unsupported or minimally supported nodes, mostly near the root of the tree, are a possible cause for the skewed mean value. A total of 439 nodes out of 4488 (9.78%) have a bootstrap value of 10 or less, with 134 nodes having 0 support.

```
tree <- read.tree(file.path
                  ('..', 'results', 'trees', 'combined_tree', 'combined_tree.tree'))
```

```r
tree$node.label <- as.numeric(tree$node.label)

summary(tree$node.label) # summary of node support values

head(data.frame(table(tree$node.label), row.names = NULL), n = 11)
# a count of the 11 lowest bootstrap values
```

Table 5: A summary of node support values

Min.	0.000
1st Qu.	38.000
Median	70.000
Mean	63.804
3rd Qu.	96.000
Max.	100.000
NA's	1.000

Table 6: A count of the 11 lowest bootstrap values

Var1	Freq
0	134
1	54
2	40
3	21
4	23
5	33
6	43
7	23
8	20
9	22
10	26