Creating file for NeighborNet Treebuilding

Grace Saville

01/10/2021

Preamble

```
library(tidyr)
library(ape)
library(phylotools)
getwd()
setwd("C:/Users/airhe/OneDrive/Documents/Masters/Project 3/kiore-project")
```

Loading the data

```
data <- read.csv("./data/ratsSNPs_clean.csv")
dim(data)   # 370 rows 298 columns
str(data)
```

Testing the dat2phylip function

```
x <- data.frame(a = c("asdfghjkl1", "asdf2ghjkl", "asdf3ghjkl", "asdfghjkl4"),
    b = c("CTAGTGACCCGTAG", "TGACCCGTTAGAAC", "TGACTCTTTAGAAC", "TTTCACGTTGAGAC"))  # dummy df
dat2phylip(x, outfile = "test.phy")  # testing this function that saves phylip files
```

After some experimenting, for Phylip files. . .

- spaces in the base strings aren't necessary
- 10 character length names are needed
- dashes "-" are allowed
- no differentiation between upper and lower case
- only letters from bases or amino acids allowed (?, e.g acdefghiklmnprstvwyz)

Replacing the SNP symbols with IUPAC Ambiguity code symbols

IUPAC Ambiguity code:

Symbol	SNP bases
A	AA
T	TT
C	CC
G	GG
R (purine)	AG
Y (pyrimidine)	CT
W (weak)	AT
S (strong)	CG
M (amino)	AC
K (keto)	GT

```r
copy <- data   # making a copy

copy[copy == "A:A"] <- "A"
copy[copy == "T:T"] <- "T"
copy[copy == "C:C"] <- "C"
copy[copy == "G:G"] <- "G"
copy[copy == "A:G"] <- "R"
copy[copy == "G:A"] <- "R"
copy[copy == "C:T"] <- "Y"
copy[copy == "T:C"] <- "Y"
copy[copy == "A:T"] <- "W"
copy[copy == "T:A"] <- "W"
copy[copy == "C:G"] <- "S"
copy[copy == "G:C"] <- "S"
copy[copy == "A:C"] <- "M"
copy[copy == "C:A"] <- "M"
copy[copy == "G:T"] <- "K"
copy[copy == "T:G"] <- "K"
```

Adjusting the data frame for Phylip output function

```r
names(copy)
copy <- copy[, -c(2:16)]   # removing the rows with information other than the species key and codes
copy <- unite(copy, bases, -1, sep = "", remove = TRUE)   # merging all the base columns into one
```

```r
dat2phylip(copy, outfile = "./data/ratsSNPs_NeighborNet_input.phy")
# can be opened by splitstree when configured as proteins/amino acids
```