Modification of the unexpected behaviour of the Character Difference

Thomas Guillerme

2019-04-18

The version of the Character Difference metric proposed in the manuscript version USYB-2018-099 had a flawed behaviour which did not capture correctly the difference between characters as defined by the difference in phylogenetic splits implied by the two characters.

This is the old version of the Character Difference metric (in manuscript USYB-2018-099 and `dispRity` package `v1.0.0`):

- Old version of the Character Difference metric:

$$CD.old_{(x,y)} = 1 - 2\left(\frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2}\right) \tag{1}$$

This version was problematic as the differences where not linear creating the maximum difference when only half of the characters where different (see below). It has now been corrected in this newer version (current manuscript and `dispRity` package >= `v1.1.0`) that is linear.

- New version of the Character Difference metric

$$CD.new_{(x,y)} = \frac{\sum_i^n |x_i - y_i|}{n - 1} \tag{2}$$

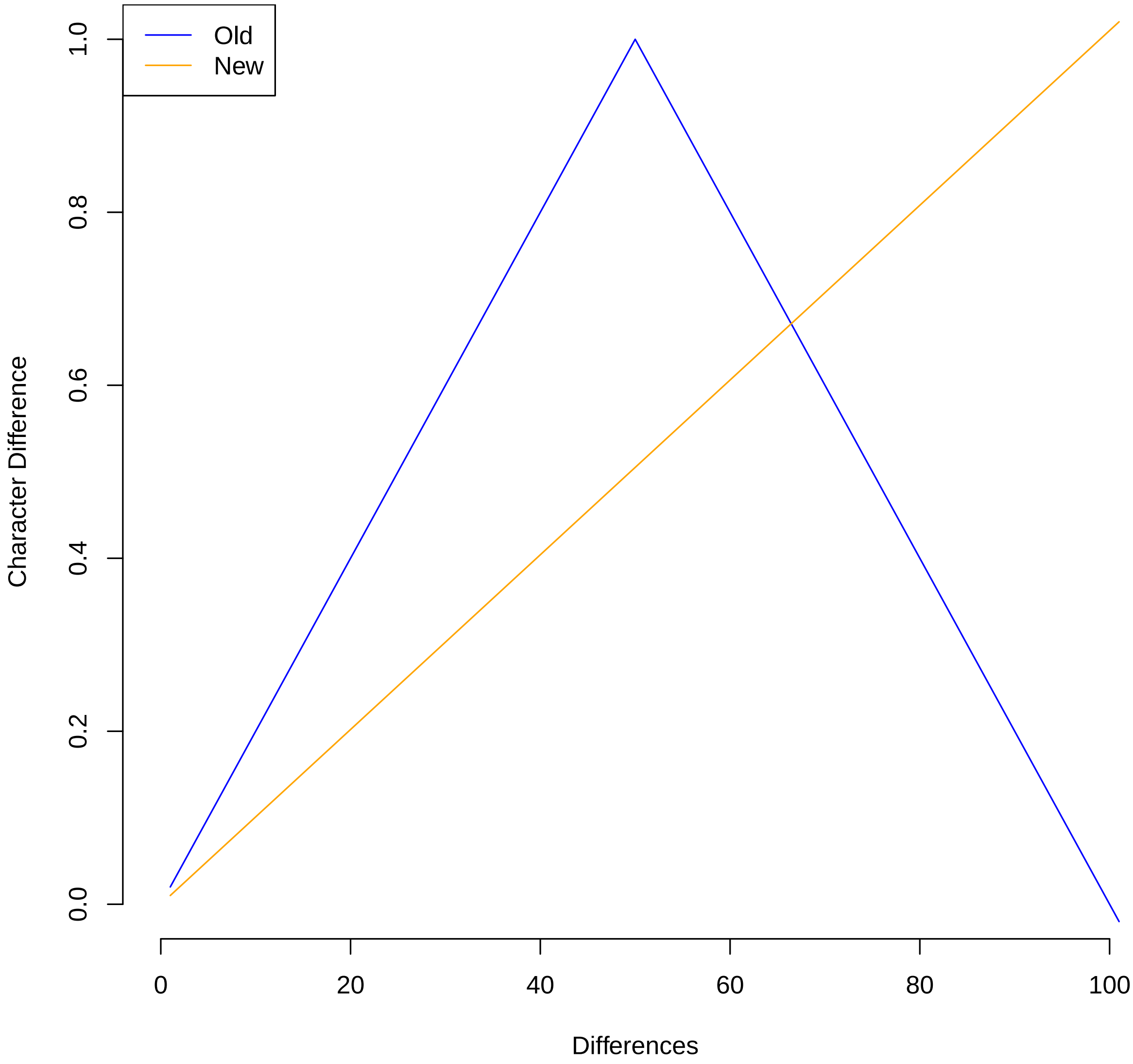
Behaviour of both metric versions

If we consider an hypothetical discrete character matrix of an unknown number of taxa and 100 characters, the difference between characters can range from 0 (all characters are identical) to 100 (every character are totally different). Considering this, we can compare both metrics considering a distance vector (`dist`) ranging from 0 to 100 and plot the Character difference for each differences:

```
## Older version
old.char.diff <- function(distance, count) {
    ## The distance
    difference <- distance/count
    ## The Character Difference
    return( 1 - ( abs(difference-0.5)/0.5 ))
}
## New version
new.char.diff <- function(distance, count) {
    ## The distance (note the (count-1))
    difference <- distance/(count-1)
    ## The Character Difference (equal to the distance / (count - 1))
    return(difference)
}

## The distances to consider
distance <- seq(from = 0:100)
```

```r
## The change in Character difference metric
par(bty = "n")
plot(NULL, xlim = c(0, 100), ylim = c(0, 1), ylab = "Character Difference", xlab = "Differences")
lines(old.char.diff(distance, count = 100), col = "blue")
lines(new.char.diff(distance, count = 100), col = "orange")
legend("topleft", c("Old", "New"), col = c("blue", "orange"), lty = 1)
```



Note that the number of taxa is unknown since in practice, a matrix that will satisfy this assumption above is highly hypothetical and is unlikely to be used in practice (i.e. that will require a matrix with more taxa than character which is rarely used in evolutionary biology).