Influence of different modes of morphological character correlation on phylogenetic tree inference

Supplementary material 1 - Character difference

Thomas Guillerme[1,2,*], and Martin D. Brazeau[2,3]

[1] *School of Biological Sciences, University of Queensland, St. Lucia, Queensland, Australia.*

[2] *Imperial College London, Silwood Park Campus, Department of Life Sciences, Buckhurst Road, Ascot SL5 7PY, United Kingdom.*

[3] *Department of Earth Sciences, Natural History Museum, Cromwell Road, London, SW75BD, United Kingdom.*

***Corresponding author.** *guillert@tcd.ie*

Definition

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

Where n is the number of taxa with comparable data and x_i and y_i are each characters states for the characters x and y and the taxa i.

Distance metric proof

The Character Difference (CD) metric is a distance $d : CD \times CD \to [0; \infty)$ with any character x, y and $z \in CD$ if an only if the followings are satisfied:

$$CD_{(x,y)} \geq 0 \tag{2}$$

$$CD_{(x,y)} = 0 \Leftrightarrow x = y \tag{3}$$

$$CD_{x,y} = CD_{y,x} \tag{4}$$

$$CD_{(x,y)} \leq CD_{(x,z)} + CD_{(z,y)} \tag{5}$$

Character difference positivity.—

For any $n \in [1;\infty)$ and x_n and $y_n \in (0,1)$ because the difference between character states is considered as Fitch-like (unordered; i.e. the difference between two different characters states is equal to 1 and the difference between two identical character states is equal to 0):

$$
\begin{gathered}
0 \leq |x_n - y_n| \leq 1 \\
0 \leq \sum_{i=1}^{n} |x_i - y_i| \leq n \\
0 \leq \frac{\sum_i^n |x_i - y_i|}{n} \leq 1 \\
0 \leq \left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| \leq \frac{1}{2} \\
0 \leq 2 \left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| \leq 1 \\
0 \leq 1 - 2 \left(\left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| \right) \leq 1
\end{gathered}
\tag{6}
$$

Thus equation 2 is true.

Character difference identity.—

For any $n \in [1;\infty)$ and $x_n = y_n$ then:

$$\begin{aligned} |x_n - y_n| &= 0 \\ \sum_i^n |x_i - y_i| &= 0 \\ \frac{\sum_i^n |x_i - y_i|}{n} &= 0 \\ \left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| &= \frac{1}{2} \\ 2\left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| &= 1 \\ 1 - 2\left(\left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| \right) &= 0 \end{aligned} \tag{7}$$

Thus equation 3 is true.

Character difference equality.—

For any $n \in [1;\infty)$ and x_n and $y_n \in (0,1)$, similarly to equation 6:

$$\begin{aligned} |x_n - y_n| &= |y_n - x_n| \\ \sum_i^n |x_i - y_i| &= \sum_i^n |y_i - x_i| \\ \frac{\sum_i^n |x_i - y_i|}{n} &= \frac{\sum_i^n |y_i - x_i|}{n} \\ \left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| &= \left| \frac{\sum_i^n |y_i - x_i|}{n} - \frac{1}{2} \right| \\ 2\left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| &= 2\left| \frac{\sum_i^n |y_i - x_i|}{n} - \frac{1}{2} \right| \\ 1 - 2\left(\left| \frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2} \right| \right) &= 1 - 2\left(\left| \frac{\sum_i^n |y_i - x_i|}{n} - \frac{1}{2} \right| \right) \end{aligned} \tag{8}$$

Thus equation 4 is true.

Character difference subadditivity.—

For any $n, j, b \in [1;\infty)$ and x_n, y_j, $z_b \in (0,1)$:

$$CD_{x,y} \leq CD_{x,z} + CD_{z,y}$$

if:

$$1-2\left(\left|\frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2}\right|\right) \leq 1-2\left(\left|\frac{\sum_k^j |x_k - z_k|}{j} - \frac{1}{2}\right|\right) + 1-2\left(\left|\frac{\sum_a^b |z_a - y_a|}{b} - \frac{1}{2}\right|\right)$$

then, the opposite is:

$$\underbrace{\left|\frac{\sum_i^n |x_i - y_i|}{n} - \frac{1}{2}\right|}_{\in[0,1/2]} \geq \underbrace{\left|\frac{\sum_k^j |x_k - z_k|}{j} - \frac{1}{2}\right|}_{\in[0,1/2]} + \underbrace{\left|\frac{\sum_a^b |z_a - y_a|}{b} - \frac{1}{2}\right|}_{\in[0,1/2]} - 2$$

thus:

$$\begin{aligned} &\in [0, \tfrac{1}{2}] \geq \in [0, \tfrac{1}{2}] + \in [0, \tfrac{1}{2}] - 2 \\ &\in [0, \tfrac{1}{2}] \geq \in [-2, -1] \end{aligned} \tag{9}$$

Thus equation 5 is true.