HOPACH - Added Distances

1 Binary distance - general form

Let $X = (x_1, x_2, \ldots, x_n)$, $Y = (y_1, y_2, \ldots, y_n)$, with $x_i, y_i \in \{0, 1\}$. Define
$N_{11} = \sum_{i=1}^{n} I(x_i = 1, y_i = 1)$,
$N_{10} = \sum_{i=1}^{n} I(x_i = 1, y_i = 0)$,
$N_{01} = \sum_{i=1}^{n} I(x_i = 0, y_i = 1)$,
$N_{00} = \sum_{i=1}^{n} I(x_i = 0, y_i = 0)$,
$N_{11} + N_{10} + N_{01} + N_{00} = n$ (see Table 1).

		Y	
		0	1
X	0	N_{00}	N_{01}
	1	N_{10}	N_{11}

Table 1

Then the distance between X and Y is

$$d(X, Y) = \frac{N_{10} + N_{01}}{\alpha_1 N_{11} + \alpha_2 N_{00} + N_{10} + N_{01}}, \tag{1}$$

where $\alpha_1 \geqslant 0,\ \alpha_2 \geqslant 0$ are tuning parameters.

Special cases:

- **Manhattan (binary)**: $\alpha_1 = 1,\ \alpha_2 = 1$
- **Jaccard**: $\alpha_1 = 1,\ \alpha_2 = 0$

2 Special binary: metametric

$d(X, Y) = 0$ implies $X = Y$ but $X = Y$ does not imply zero distance.

$$d(X, Y) = \frac{N_{10} + N_{01} + 0.5\, N_{00}}{n} \tag{2}$$

3 Continuous version (S-function distance)

This metric is designed for use with p-values, where x_i's and y_i's are p-values used to determine which elements of X and Y contain some kind of significant response (e.g. differential expression). For this situation, binary distances can be used as well, encoding significance as 1 and non-significance as 0. Continuous version offers a smooth curve in place of a step function of a hard cut-off, which might be undesirable since cut-offs are in general arbitrary, and adjusted p-values are affected by various decisions such as a choice of a multiple testing correction.

This distance can be viewed as a continuous generalization of a Jaccard distance approach - removal of (x_i, y_i) for $\{i : x_i > 0.2, y_i > 0.2\}$ ("non-significant" positions, analog of 00 in binary distance) followed by normalized Manhattan distance of transformed data on remaining positions:

$$d(X, Y) = \frac{\sum_{i=1}^{n} \Big[1 - I(x_i > 0.2) I(y_i > 0.2)\Big] \, |f(x_i) - f(y_i)|}{\sum_{i=1}^{n} \Big[1 - I(x_i > 0.2) I(y_i > 0.2)\Big]}, \tag{3}$$

where

$$f(x) = 1 - e^{-a\, x^b}. \tag{4}$$

The transformation function $f(x)$ has an asymmetric inverted sigmoid shape (S-function). It descends quickly to 0 on the left side for small p-values and allows for a more gradual ascent to 1 on the right (Figure 1). The shape of the curve can be adjusted with tuning parameters a and b.

4 Note

For all the metrics above, $0 \leqslant d(X, Y) \leqslant 1$.

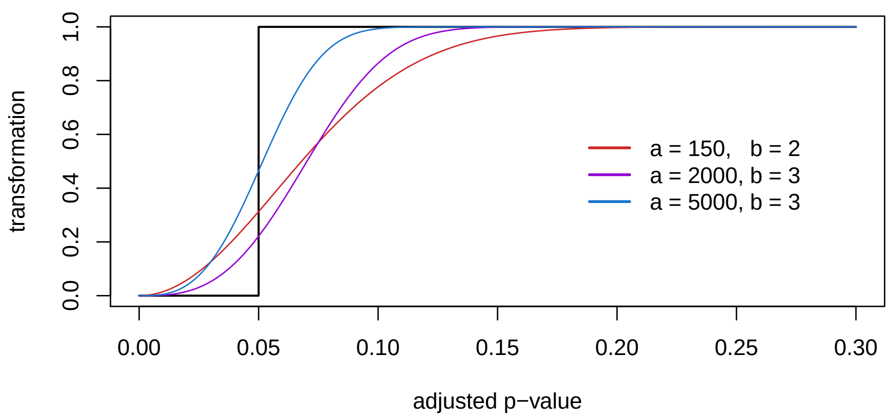

Figure 1: Transformation function for the p-value-based distance