A Bayesian Model of the DNA Barcode Gap

Supplementary Information

Jarrett D. Phillips

1 Mathematical Details

For a given random variable X, its cumulative distribution function (CDF) is defined by

$$F_X(t) = \mathbb{P}(X \leq t). = 1 - \mathbb{P}(X > t) \tag{1}$$

Equations (1)-(4) within the main text can be expressed in terms of empirical cumulative distribution functions (ECDFs) as follows, since the true underlying CDFs, $F(\cdot)$, are unknown *a priori*:

$$\begin{aligned} p_x &= \mathbb{P}(d_{ij} \geq a) \\ &= 1 - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) \end{aligned} \tag{2}$$

$$\begin{aligned} q_x &= \mathbb{P}(d_{XY} \leq b) \\ &= \hat{F}_{d_{XY}}(b) \end{aligned} \tag{3}$$

$$\begin{aligned} p'_x &= \mathbb{P}(d_{ij} \geq a') + \mathbb{P}(d_{XY} = a') \\ &= 1 - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{XY} = a') \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{XY} = a') \end{aligned} \tag{4}$$

$$\begin{aligned} q'_x &= \mathbb{P}(d'_{XY} \leq b) \\ &= \hat{F}_{d'_{XY}}(b) \end{aligned} \tag{5}$$

noting that $\mathbb{P}(d_{ij} \leq b) = \hat{F}_{d_{ij}}(b) = 1$, $\mathbb{P}(d_{XY} \leq a) = \hat{F}_{d_{XY}}(a) = 0$, and $\mathbb{P}(d'_{XY} \leq a') = \hat{F}_{d'_{XY}}(a') = 0$ (see **Figure 1** in main text). Given n increasing-ordered data points, the (discrete) ECDF, $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[x_i \leq t]}$, comprises a step function having jump discontinuities of size $\frac{1}{n}$ at each sample observation (x_i), excluding ties (or steps of weight $\frac{i}{n}$ with tied observations), where $\mathbb{1}(x)$ is the indicator function. Equations (1)-(4) herein

clearly demonstrate the asymmetric directionality of the proposed metrics. Furthermore, calculation of the DNA barcode gap estimators given in the main text is straightforward as they implicitly account for both total distribution area and overlap. Nevertheless, the total areas bounded by intraspecific and interspecific distributions on $[a,\, b]$, and combined distributions on $[a',\, b]$ (see **Figure 1** in main text) are given by the joint ECDFs, $\hat{F}_{d_{ij},\, d_{XY}}(a, b) = \mathbb{P}(d_{ij} \geq a, d_{XY} \leq b)$ and $\hat{F}_{d_{ij},\, d'_{XY}}(a', b) = \mathbb{P}(d_{ij} \geq a', d'_{XY} \leq b)$. Thus,

$$\begin{aligned} A &= p_x + q_x \\ &= 1 - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) + \hat{F}_{d_{XY}}(b) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) + \hat{F}_{d_{XY}}(b) \qquad (6) \\ A' &= p'_x + q'_x \\ &= 1 - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') + \hat{F}_{d'_{XY}}(b) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') + \hat{F}_{d'_{XY}}(b) \qquad (7) \end{aligned}$$

whose values lie in [0, 2]. Similarly, the degree of distribution overlap is

$$\begin{aligned} O &= \hat{F}_{\min(d_{ij},\, d_{XY})}(a, b) \\ &= \hat{F}_{\min(d_{ij},\, d_{XY})}(b) - \hat{F}_{\min(d_{ij},\, d_{XY})}(a) \qquad (8) \\ O' &= \hat{F}_{\min(d_{ij},\, d'_{XY})}(a', b) \\ &= \hat{F}_{\min(d_{ij},\, d'_{XY})}(b) - \hat{F}_{\min(d_{ij},\, d'_{XY})}(a') \qquad (9) \end{aligned}$$

whose support is on [0, 1]. All above estimators are (uniformly and pointwise) consistent, approaching their true (function) values almost surely, and therefore also in probability to one, as the sample size tends to infinity, as guaranteed by both the Strong and Weak Laws of Large Numbers, and by the Glivenko-Cantelli Theorem.