A Bayesian Model of the DNA Barcode Gap

Jarrett D. Phillips[1,3]* (ORCID: 0000-0001-8390-386X)

[1]*School of Computer Science, University of Guelph, Guelph, ON., Canada, N1G2W1*

***Corresponding Author**: Jarrett D. Phillips[1]

Email Address: jphill01@uoguelph.ca

Running Title:

Abstract

Keywords: Bayesian inference, DNA barcoding, intraspecific genetic distance, interspecific genetic distance, specimen identification, species discovery, Stan

1 Introduction

Since its inception over 20 years ago, DNA barcoding (Hebert et al., 2003a,b) has emerged as a robust method of specimen identification and species delimitation across myriad taxonomic groups which have been sequenced at short, standardized gene regions like 5'-COI for animals. However, the success of the approach depends crucially on two important factors: (1) the availability of high-quality specimen records found in public reference sequence databases such as the Barcode of Life Data Systems (BOLD) (Ratnasingham and Hebert, 2007), and (2) the establishment of a DNA barcode gap — the idea that the maximum genetic distance observed within species is much smaller than the minimum degree of marker variation found among species (Meyer and Paulay, 2005; Meier et al., 2008). Early work has demonstrated that the presence of a DNA barcode gap hinges strongly on extant levels of species haplotype diversity gauged from comprehensive specimen sampling at wide geographic and ecological scales. Despite this, many taxa lack adequate separation in their pairwise intraspecific and interspecific genetic distances, thereby compromising rapid matching of unknown samples to expertly-validated references.

Recent work has argued that DNA barcoding, in its current form, is lacking in statistical rigor, calling into question the existence of a true species' DNA barcode gap (Phillips et al., 2022). To support this notion, novel nonparametric locus-specific metrics based on the multispecies coalescent (Rannala and Yang, 2003; Yang and Rannala, 2017) were recently outlined and shown to hold strong promise when applied to predatory *Agabus* (Coleoptera: Dytiscidae) diving beetles (Phillips et al., 2024). The coalescent (Kingman, 1982) encompasses a backwards continuous-time stochastic Markov process of allelic sampling within natural, neutrally-evolving, species populations towards the Most Recent Common Ancestor (MRCA). The metrics quantify the extent of asymmetric directionality of proportional genetic distance distribution overlap/separation for species within well-sampled taxonomic genera based on a straightforward distance count. The metrics can be employed in a variety of ways, including to assess performance of marker genes for species identication, as well as to assess whether computed values are consistent with population genetic-level parameters like effective population size (N_e), mutation rates (μ) and divergence times (τ) for species under study (Mather et al., 2019). However, what appears to be missing is an unbiased way to compute the statistical accuracy of the recommended estimators arising through problems inherent in frequentist maximum likelihood estimation for discrete probability distributions having bounded positive support on [0, 1]. To this end, here, a Bayesian model of the DNA barcode gap coalescent is introduced to rectify such issues. The model allows accurate estimation of posterior means, posterior standard deviations, posterior quantiles, and credible intervals for the metrics given datasets of intraspecific and interspecific genetic distances for species of interest.

2 Methods

2.1 DNA Barcode Gap Metrics

Recently, Phillips et al. (2024) proposed novel nonparametric maximum likelihood estimators (MLEs) of proportional overlap/separation between intraspecific and interspecific pairwise genetic distance distributions for a given species (x) to aid assessment of the DNA barcode gap as follows:

$$p_x = \frac{\#\{d_{ij} \geq a\}}{\#\{d_{ij}\}} \tag{1}$$

$$q_x = \frac{\#\{d_{XY} \leq b\}}{\#\{d_{XY}\}} \tag{2}$$

where d_{ij} and d_{XY} are distances within and among species, respectively, and the notation # reflects a count (**Figure 1**). Quantities a and b correspond to $\min(d_{XY})$ and $\max(d_{ij})$, the minimum interspecific distance and the maximum intraspecific distance, respectively. Notice that a and b are also the first and nth order statistics, respectively. Distances are easily computed from a model of DNA sequence evolution, such as p distance. Similar expressions (denoted $p_x^{'}$ and $q_x^{'}$) for nearest neighbour species were also given (see Phillips et al. (2024)), in which d_{XY} included only interspecific distances between the species of interest and its closest neighbouring species. If a focal species is found to have multiple nearest neighbours, then the species possessing the smallest average pairwise interspecfic distance is used. While these schemes differ considerably from the usual definition of the DNA barcode gap laid out by Meyer and Paulay (2005) and Meier et al. (2008), they more accurately account for species' coalescence histories inferred from contemporaneous samples of DNA sequences. such as interspecific hybridization/introgression events (Phillips et al., 2024). Note, distances (and hence the metrics) are constrained to the closed interval [0, 1]. Values of the estimators obtained from equations (1) and (2) close to or equal to zero give evidence for separation between intraspecific and interspecific genetic distance distributions; that is, values suggest the presence of a DNA barcode gap for a target species. Conversely, values near or equal to one give evidence for distribution overlap; that is, values likely indicate the absence of a gap. Equations (1) and (2) can be expressed in terms of empirical cumulative distribution functions (ECDFs)

$$p_x = \mathbb{P}(d_{ij} \geq a) \quad = 1 - \hat{F}_{d_{ij}}(a) = \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) \tag{3}$$

$$q_x = \mathbb{P}(d_{XY} \leq b) = \hat{F}_{d_{XY}}(b), \tag{4}$$

noting that $\hat{F}_{d_{XY}}(a) = 0$ (**Figure 1**). Given n increasing-ordered data points, the 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 $\frac{1}{n}$ at each sample observation (x_i), excluding ties, where $\mathbb{1}(x)$ is the indicator function. From here, the asymmetric directionality of the metrics is obvious. As mentioned previously, similar equations for $p_x^{'}$ and $q_x^{'}$ can be easily derived.

2.2 A Bayesian Implementation

A major criticism of large sample (frequentist) theory is that it relies on asymptotic properties of the MLE (which is assumed to be a fixed but unknown quantity), such as estimator normality and consistency. This problem is especially pronounced in the case of binomial proportions. The estimated Wald SE of the sample proportion, is given by

$$\widehat{SE[\hat{p}]} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{5}$$

where $\hat{p} = \frac{Y}{n}$ is the MLE, Y is the number of successes ($Y = \sum_{i=1}^{n} y_i$) and n is the number of trials (*i.e.*, sample size). However, the above formula is problematic for several reasons. First, Equation (5) makes use of the Central Limit Theorem (CLT); thus, large sample sizes are required for reliable estimation. When few observations are available, SEs will be large and inaccurate, leading to low statistical power. Further, resulting interval estimates could span values less than zero or greater than one, or have zero width, which is practically meaningless. Second, when proprtions are exactly equal to zero or one, resulting SEs will be exactly zero, rendering Equation (5) completely useless. In the context of the proposed DNA barcode gap metrics, values obtained at the boundaries of their support are often encountered. Therefore, reliable calculation of SEs is not feasible. Given the importance of sufficient sampling of species genetic diversity for DNA barcoding initiatives, a different statistical estimation approach is necessary. Bayesian inference offers a natural path forward in this regard since it allows for straightforward specification of prior beliefs concerning unknown model parameters and permits the seamless propagation of uncertainty, when data is lacking, through integration with the likelihood function associated with true generating processes. As a consequence, Bayesian models are much more flexible and generally more easily interpretable compared to frequentist approaches since entire posterior distributions, along with their summaries, are outputted, rather than just sampling distributions, p-values, and confidence intervals, allowing direct probability statements to be made.

2.3 The Model

Essentially, from a statistical perspective, the goal herein is to nonparametrically estimate probabilities corresponding to extreme tail quantiles for positive highly skewed distributions on the unit interval. Here, it is sought to numerically approximate the extent of overlap/separation of intraspecific and interspecific pairwise genetic distance distributions within $[a, b]$. This is a challenging computational problem within the current study as detailed in subsequent sections. Counts, y, of overlapping genetic distances (as expressed in the numerator of Equations (1) and (2)) are treated as binomially distributed with expectation $\mathbb{E}[Y] = k\theta$, where $k = \{N, M, C\}$ are total counts of intraspecific, interspecific, and combined genetic distances for a target species along with its nearest neighbour species, and $\theta = \{p_x, q_x, p_x^{'}, q_x^{'}\}$. The metrics encompassing θ are presumed to follow a beta(α, β) distribution, with real shape parameters α and β, which is a natural choice of prior on probabilities. Such a scheme is quite convenient since the beta distribution is conjugate to the binomial distribution. Thus, the posterior distribution is also beta distributed. Parameters were given an uninformative Beta(1, 1) prior, which is equivalent to a standard uniform (Uniform(0, 1)) prior since it places equal probability on all parameter values within its support. As a result, the posterior is Beta$(Y+1,\, n-Y+1)$, from which various moments and other quantities, such as the expected value $\mathbb{E}[Y] = \frac{Y+1}{n+2}$ and variance $\mathbb{V}[Y] = \frac{(Y+1)(n-Y+1)}{(n+2)^2(n+3)}$, can be easily calculated. In general however, when possible, it is always advisable to incorporate prior information, even if only weak, rather than simply imposing complete ignorance in the form of a flat prior distribution. With sufficient data, the choice of prior distribution becomes less important since the posterior will be directly proportional to the likelihood. The full univariate Bayesian model for species x is thus given by

$$\begin{aligned} y_{\text{lwr}}[x] &\sim \text{Binomial}(N[x], p_{\text{lwr}}[x]) \\ y_{\text{upr}}[x] &\sim \text{Binomial}(M, p_{\text{upr}}[x]) \\ y_{\text{lwr}}^{'}[x] &\sim \text{Binomial}(N[x], p_{\text{lwr}}^{'}[x]) \\ y_{\text{upr}}^{'}[x] &\sim \text{Binomial}(C[x], p_{\text{upr}}^{'}[x]) \\ p_{\text{lwr}}[x], p_{\text{upr}}[x], p_{\text{lwr}}^{'}[x], p_{\text{upr}}^{'}[x] &\sim \text{Beta}(1,1). \end{aligned} \tag{6}$$

The model was fitted using the Stan probabilistic programming language (Carpenter et al., 2017) framework for Hamiltonian Monte Carlo (HMC) via the No-U-Turn Sampler (NUTS) sampling algorithm (Hoffman and Gelman, 2014) through the `rstan` R package (Stan Development Team, 2023). Four chains were run for 2000 iterations each in parallel across four cores with random paramester initializations. Within each chain, a total of 1000 samples was discarded as warmup (*i.e.*, burnin) to reduce dependence on starting conditions. Further, 1000 post-warmup draws were utilized per chain. Because HMC/NUTS results in dependent samples that are minimally autocorrelated, Markov chain thinning is not required. Each of these reflect default MCMC settings in Stan. Since the DNA barcode gap metrics often attain values very close to zero and/or very near one, in addition to more intermediate values, a Beta$(\frac{1}{2}, \frac{1}{2})$ prior, which is U-shaped symmetric and places greater probability density at the extremes of the distribution due to its heavier tails, while still allowing for variability in parameter estimates within intermediate values along its domain, was also attempted. However, this resulted in several divergent transitions, among other pathologies, imposed by complex geometry (*i.e.*, curvature) in the posterior space, despite remedies to resolve them, such as lowering the step size of the HMC sampler. Note that this prior is Jeffreys' prior, which is proportional to the square root of the Fisher information and has several desirable statistical properties, most notably invariance to reparameterization.

To validate the overall correctness of the proposed statistical model, in addition to generating MLEs as a means of comparison, posterior predictive checks were also employed to generate binomial random variates in the form of counts from the posterior predictive distribution; that is $\gamma = \{Np_x, Mq_x, Np_x^{'}, Cq_x^{'}\}$ to verify that the model adequately captures relevant features of the observed data.

2.4 Case Study

To demonstrate the promise of the proposed Bayesian estimation approach, the model has been written to be applied to several species within a well-sampled genus of interest. Specifically,

3 Results

4 Discussion

5 Conclusion

Supplementary Information

Information accompanying this article can be found in Supplemental Information.pdf.

Data Availability Statement

Raw data, R, and Stan code can be found on GitHub at: https://github.com/jphill01/Bayesian-DNA-Barcode-Gap-Coalescent.

Acknowledgements

We wish to recognise the valuable comments and discussions of Daniel (Dan) Gillis, Robert (Bob) Hanner, Robert (Rob) Young, and XXX anonymous reviewers.

We acknowledge that the University of Guelph resides on the ancestral lands of the Attawandaron people and the treaty lands and territory of the Mississaugas of the Credit. We recognize the significance of the Dish with One Spoon Covenant to this land and offer our respect to our Anishinaabe, Haudenosaunee and Métis neighbours as we strive to strengthen our relationships with them.

Funding

None declared.

Conflict of Interest

None declared.

Author Contributions

JDP wrote the manuscript, wrote R and Stan code, approved all developed code as well as analysed and interpreted all experimental results.

References

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell
2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1.

Hebert, P., A. Cywinska, S. Ball, and J. deWaard
2003a. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512):313–321.

Hebert, P., S. Ratnasingham, and J. de Waard
2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 1):S96–S99.

Hoffman, M. and A. Gelman
2014. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Kingman, J.
1982. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248.

Mather, N., S. M. Traves, and S. Y. W. Ho
2019. A practical introduction to sequentially markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, 10(1):579–589.

Meier, R., G. Zhang, and F. Ali

2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic Biology*, 57(5):809–813.

Meyer, C. and G. Paulay

2005. DNA barcoding: error rates based on comprehensive sampling. *PLOS Biology*, 3(12):e422.

Phillips, J., D. Gillis, and R. Hanner

2022. Lack of statistical rigor in DNA barcoding likely invalidates the presence of a true species' barcode gap. *Frontiers in Ecology and Evolution*, 10:859099.

Phillips, J., C. Griswold, R. Young, N. Hubert, and H. Hanner

2024. *A Measure of the DNA Barcode Gap for Applied and Basic Research*, Pp. 375–390. New York, NY: Springer US.

Rannala, B. and Z. Yang

2003. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164:1645–1656.

Ratnasingham, S. and P. Hebert

2007. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, 7(3):355–364.

Stan Development Team

2023. RStan: The R interface to Stan. R package version 2.21.8.

Yang, Z. and B. Rannala

2017. Bayesian species identification under the multispecies coalescent provides significant improvements to dna barcoding analyses. *Molecular Ecology*, 26:3028–3036.

Figures

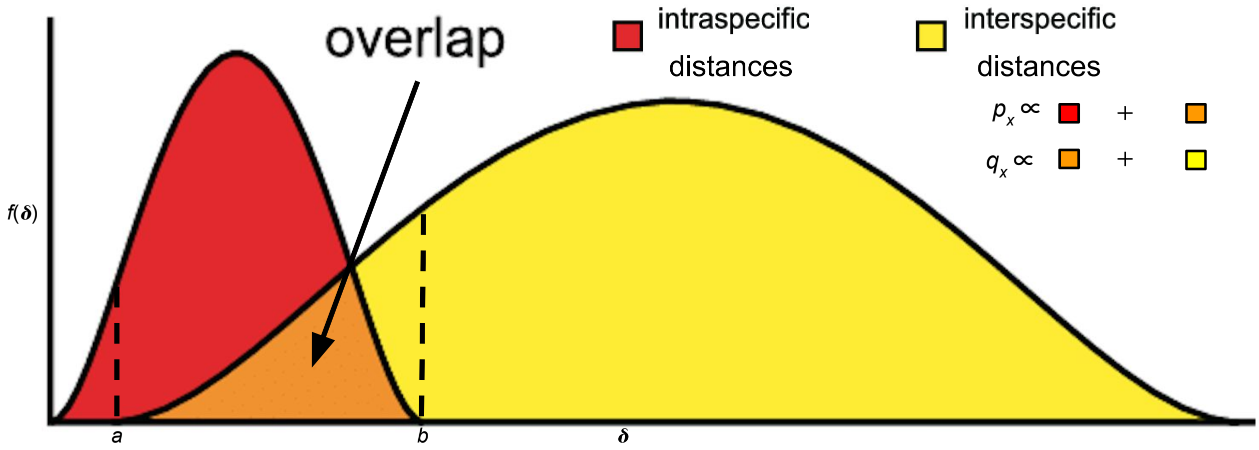

Figure 1: Modified depiction from Meyer and Paulay (2005) and Phillips et al. (2024) of the overlap/separation of pairwise intraspecific and interspecific genetic distances (δ) for calculation of the DNA barcode gap metrics (p_x and q_x) for species x. The minimum interspecific distance is denoted by a and the maximum intraspecific distance is indicated by b. The quantity $f(\delta)$ is akin to a kernel density estimate of the probability density function of pairwise genetic distances. A similar visualization can be displayed for $p_x^{\prime}$ and $q_x^{\prime}$.