

Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture


2009 - 21st Annual Conference Proceedings

SEQUENTIAL BAYESIAN CLASSIFICATION: DNA BARCODES

Michael P. Anderson

Suzanne Dubnicka

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

 Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Anderson, Michael P. and Dubnicka, Suzanne (2009). "SEQUENTIAL BAYESIAN CLASSIFICATION: DNA BARCODES," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1083>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

SEQUENTIAL BAYESIAN CLASSIFICATION: DNA BARCODES

Michael P. Anderson and Suzanne Dubnicka

Department of Statistics, Kansas State University, Manhattan, KS 66506-0803

Abstract

DNA barcodes are short strands of nucleotide bases taken from the cytochrome c oxidase subunit 1 (COI) of the mitochondrial DNA (mtDNA). A single barcode may have the form C C G G C A T A G T A G G C A C T G and typically ranges in length from 255 to around 700 nucleotide bases. Unlike nuclear DNA (nDNA), mtDNA remains largely unchanged as it is passed from mother to offspring. It has been proposed that these barcodes may be used as a method of differentiating between biological species (Hebert, Ratnasingham, and deWaard 2003). While this proposal is sharply debated among some taxonomists (Will and Rubinoff 2004), it has gained much momentum and attention from biologists. One issue at the heart of the controversy is the use of genetic distance measures as a tool for species differentiation. Current methods of species classification utilize these distance measures that are heavily dependent on both evolutionary model assumptions as well as a clearly defined “gap” between intra- and interspecies variation (Meyer and Paulay 2005). We point out the limitations of such distance measures and propose a character-based method of species classification which utilizes an application of Bayes’ rule to overcome these deficiencies. The proposed method is shown to provide accurate species-level classification. The proposed methods also provide answers to important questions not addressable with current methods.

Keywords: DNA barcoding, Bayesian methods, sequential analysis, classification

1 Introduction

1.1 Challenges in Taxonomy

Taxonomists face great challenges regarding the classification and discovery of congeneric, or closely related, species. In order to determine an organism’s species, taxonomy relies mainly upon inspection of an organisms easily observed and described morphologic features, such as shapes, colors, sizes, and behaviors. These morphologic features are then compared against what is known about species that have been previously observed and classification follows.

Reliance upon these physical characteristics, or morphological features, to determine species proves challenging for several reasons. First, physical characteristics between two congeneric species may be so similar that they are mistakenly categorized as the same species. On the other hand, physical characteristics between two organisms of the same species may appear quite different, resulting in the classification of two separate species. Males and females of the same species often have different physical characteristics which can make proper

classification difficult. This latter situation can also happen when morphologic features for a particular species develop as the organism ages. For example, some frog species are easily distinguished at maturation due to color features (such as spots or stripes), but as tadpoles, these features are absent making classification at that stage very difficult and much less certain.

Second, because these morphologic features can only be compared to what has already been observed, the process of discovering new species can be quite slow. According to [Hebert, Ratnasingham, and deWaard \(2003\)](#), an individual taxonomist can rarely identify more than 1000-1500 different species. This means that when observed features do not match those of any species known to a taxonomist, there must be a large amount of collaboration with others before it can be decided a new species has been discovered. We can get a feel for the magnitude of the task at hand by noting that, in the millennia of recorded history, of the estimated 10-15 million species on earth, excluding bacteria and archaea ([Hammond 1992](#)), taxonomists have classified roughly 1.7 million of them ([Stoeckle 2003](#)).

Lastly, it is sometimes necessary to make species classifications from organism fragments. These fragments may not include enough morphologic detail to assign the organism to a species with any amount of certainty. For example, a scientist may be interested in exploring the reasons why birds are sometimes attracted to, and collide with, certain types of aircraft. After a collision of a bird with an aircraft at such high speeds, a morphologically devoid fragment of the bird may be all that remains, leaving questions about the bird's species difficult to answer. Natural history museums often contain repositories of organism fragments that account for a large amount of the biodiversity on earth. A systematic method of identification for these archival organism fragments could prove to be an important step in the direction of classifying all of the species on earth.

These difficulties in classification and discovery of species, necessitate research for a more precise and speedy discrimination among species that can complement the challenges of classification based solely on morphological features. It would be desirable to develop a method of species identification that would prove effective at every stage of the organism's life, provide quick and efficient comparisons of organisms to discover new species, and allow for proper grouping of species based on less than the complete organism.

1.2 DNA Barcoding

Developments in genetic research indicate that a short DNA sequence known as a barcode, taken from the cytochrome c oxidase subunit 1 (COI) location of mitochondrial DNA, (mtDNA) is an effective marker for identifying species in the animal kingdom ([Hebert, Ratnasingham, and deWaard 2003](#)). This barcode contains a sequence of the nucleotide bases adenine (A), thymine (T), cytosine (C), and guanine (G). A single barcode may have the form C C G G C A T A G T A G G C A C T G . . . and typically ranges in length from 255 to around 700 nucleotide bases.

Using these barcodes to discriminate between species, it is hopeful that the challenges facing taxonomy mentioned in Section 1.1 can easily be addressed. To be sure, barcodes

can be retrieved from a very small amount of an organism's tissue (1-3mm³), at any stage of life, and obtaining these barcodes is a relatively quick and inexpensive procedure costing \$3 – 5 per barcode (Hajibabaei, DeWaard, Ivanova, Ratnasingham, Dooh, Kirk, Mackie, and Hebert 2005). Thus, organism fragments and development or change of morphological features over time do not represent significant obstacles for DNA barcoding.

Initial studies of intra- and interspecies variation show, that while barcodes for the same species may not be identical, they will rarely have more than 2% divergence and will often have less than 1% divergence (Johns and Avise 1998). Hebert, Ratnasingham, and deWaard (2003) found that between species that are closely related, sequence divergence averaged around 6.8%, with 99.98% having sequence divergence greater than 3% with even larger levels of divergence among species that are not closely related. These findings imply a substantial genetic gap between these two types of variation and have led to “distance-based” methods of species classification based on the disparity between a novel barcode to be assigned to a species, and a set of barcodes that serve as a reference data set.

Distance methods, such as p-distance (Hebert, Ratnasingham, and deWaard 2003), employ thresholds derived from the aforementioned initial studies, while others such as Kimura's Two Parameter (K2P) model use assumption rich, evolution based models in addition to these thresholds (Kimura 1980).

2 Current Methods of Classification

The Consortium for the Barcode of Life (CBOL) has established standard methods of barcode classification via the Barcode of Life Data System (BOLD, www.barcodinglife.org). Ratnasingham and Hebert (2007) provide a detailed overview of the BOLD system from how barcodes are stored and accessed to the classification of new barcodes. See also Kelly, Sarkar, Eernisse, and DeSalle (2006) and Frézal and Leblois (2008). Classification of a novel barcode using the BOLD system proceeds as follows. First, the basic local alignment search tool (BLAST, (Altschul 1990)) of the BOLD data base is implemented to retrieve barcodes from reference data set that have similar features with the barcode to be classified. This search returns the top 100 matches in terms of common features between the novel barcode and the barcodes in the BOLD data base. Then distance measures, such as K2P which is the default, are computed. Next, the relationship between the new barcode and the top matches is assessed by using these distance measures together with the neighbor-joining method (Saitou and Nei 1987) to reconstruct a phylogenetic tree made up of the top 100 matches and the new barcode to be classified. The new barcode is then classified as belonging to the species of its closest neighbor in the tree, regardless of the distance between them (Frézal and Leblois 2008).

While this process of classification is fast, it leaves many important questions unanswered and has several limitations. First, it is prone to high rates of false matches in that it will classify the new barcode to its closest match regardless of the genetic distance between the two (Koski and Goulding 2001). This severely limits the ability of DNA barcoding to aid in the discovery of new species. Second, the probability that the barcode actually belongs

Species	Truncated Barcode
θ_1	C C G G C A T A G T A G G C A C T G
θ_1	C C G G C A T A G T A G G C A C T G
θ_1	C C G G C A T A G T T G G C A C T G
θ_1	C T G G C A T A G T A G G T A C T G
θ_2	C C G G C A T A G T A G G A A C A G
θ_2	C T G G C A T A G T A G G A A C A G
θ_2	C C G G C A T A G T A G G A A C A G
θ_3	C C G G A A T A G T A G G T A C C G
θ_3	C C G G A A T A G T A G G T A C C G
θ_3	C C G G A A T A G T A G G T A C C G

Table 1: Truncated Barcode Data. Only the first 18 positions of the barcode are shown here. Typical barcodes range in length from 255 positions, to 690 positions.

to the species to which it was classified can only be measured by computing percentages of similarity or genetic distance, which measures lack solid probabilistic interpretation and have been shown by [Ferguson \(2002\)](#) to be somewhat unreliable. Third, the distance measures mentioned above erase all character information when distances are computed leading to a loss of information ([DeSalle 2006](#)). Even more troubling is the work of [Meyer and Paulay \(2005\)](#) which demonstrates that the supposedly well-separated genetic gap, upon which the efficacy of distance measures is predicated, may not be so well-separated when comprehensive data sets are considered. They show that using a reference data set that contains just a few observations per species (1-2 individuals) severely underestimates intra-species variation, and they further argue that there may be much more overlap between these two types of variation than was previously assumed. The accuracy and overall relevance of these classification methods, which depend on this clearly defined gap, then come into question. Finally, the current method does not provide any assessment of how much of the barcode is necessary for proper classification. A somewhat arbitrary minimum of 500 base positions per sequence is required for inclusion into the BOLD data base, but little justification as to this particular length is provided. Such large sequences from fresh mtDNA are easily obtained, but often for archival mtDNA more than a decade old, sequences of more than 300-400 base positions long are rare ([Hajibabaei, DeWaard, Ivanova, Ratnasingham, Dooh, Kirk, Mackie, and Hebert 2005](#)).

As an alternative, we propose a character-based method of classification that assigns theoretically sound probabilities to all positive classifications. This method does not heavily rely upon the thresholds induced by the genetic gap and does not make any genetic/evolutionary model assumptions. This proposed method can also provide information as to how much of the barcode is necessary for proper classification and aid in species discovery.

Position 1	Position 2	Position 3	...
$P(A \theta_1) = 0$	$P(A \theta_1) = 0$	$P(A \theta_1) = 0$...
$P(T \theta_1) = 0$	$P(T \theta_1) = \frac{1}{4}$	$P(T \theta_1) = 0$...
$P(C \theta_1) = \frac{4}{4}$	$P(C \theta_1) = \frac{3}{4}$	$P(C \theta_1) = 0$...
$P(G \theta_1) = 0$	$P(G \theta_1) = 0$	$P(G \theta_1) = \frac{4}{4}$...

Table 2: Conditional probabilities for the first 3 positions of species θ_1 .

3 Proposed Method of Classification

The method proposed here is aimed at answering the questions left open by current methods that were discussed in Section 2. It should also be noted that, while the proposed method is presented in the context of DNA barcoding, it is general enough to be applied to other situations in which the goal is to classify high dimensional data.

The proposed method computes the probability that the new barcode belongs to each of the species at each position via an application of Bayes' rule. This is done by defining θ_l to be the event that the barcode to be classified belongs to species l , where $l = 1, \dots, s$, and s is the number of species in the reference data set. Next, prior probabilities, $P(\theta_l)$, are selected for each species and conditional probabilities of the values A, T, C, and G for each species, $P(\cdot^{(j)}|\theta_l)$, are computed at each position j in the reference data set where $j = 1, \dots, p$ with p as the number of positions on the barcode. This is done by computing the proportion of observed bases at each position within each species. For example, Table 1 contains the first 18 positions of the barcodes for a set of 10 organisms from 4 different species $\theta_1, \theta_2, \theta_3$, and θ_4 . For the first three positions of species θ_1 , the conditional probabilities are given in Table 2. Once all the conditional probabilities are constructed and prior probabilities are specified for each species, posterior calculations for a new barcode are done sequentially at each position according to the following equation:

$$P(\theta_l|\cdot^{(j)}) = \frac{P(\theta_l|\cdot^{(j-1)})P(\cdot^{(j)}|\theta_l)}{\sum_{l=1}^s P(\theta_l|\cdot^{(j-1)})P(\cdot^{(j)}|\theta_l)} \quad (1)$$

where $j = 1, \dots, k$ with k as the number of positions on the barcode, $P(\theta_l|\cdot^{(j)})$ is the posterior probability that the barcode belongs to species l after observing the nucleotide at position j , and $P(\theta_l|\cdot^{(j-1)})$ is the posterior probability that the barcode belongs to species l after observing the nucleotide at position $j - 1$ and serves as the prior probability in the posterior calculation for position j . The initial value of $P(\theta_l|\cdot^{(0)})$ is equal to the specified prior $P(\theta_l)$. The posterior probability that the barcode belongs to species θ_l at the current position then becomes the prior probability for the calculation in the next position. Hence, this method provides a sequential calculation of the probability that the barcode belongs to any of the s species in the dataset.

If $x^{(1)}, \dots, x^{(p)}$ represent the observed nucleotides along a barcode sequence with p positions, and θ_l is the event the barcode belongs to species l , then the goal of the proposed method's calculation is to compute $P(\theta_l|x^{(1)}, \dots, x^{(p)})$.

Theorem 1. Let $x^{(1)}, \dots, x^{(p)}$ be p independent observations that arise in sequence. Suppose that $P(\theta_l)$ represents the prior probability that the sequence of observations belongs to group l . Suppose further that the conditional probabilities $P_1(x^{(1)}|\theta_l), \dots, P_p(x^{(p)}|\theta_l)$ are known.

Then, by using the posterior probability from position j , $P(\theta_l|x^{(1)}, \dots, x^{(j)})$ as the prior probability in equation (1) for computing the posterior at position $j+1$, $P(\theta_l|x^{(1)}, \dots, x^{(j+1)})$, sequentially for $j = 1, \dots, p$, results in computing $P(\theta_l|x^{(1)}, \dots, x^{(p)})$.

Proof. First notice that by independence of the observations $x^{(1)} \dots x^{(p)}$ we have

$$P(\theta_l|x^{(1)}, \dots, x^{(p)}) = \frac{P(\theta_l)P_1(x^{(1)}|\theta_l)P_2(x^{(2)}|\theta_l) \cdots P_p(x^{(p)}|\theta_l)}{P_1(x^{(1)})P_2(x^{(2)}) \cdots P_p(x^{(p)})} \quad (2)$$

Now using the prior probability $P(\theta_l)$, the posterior probability for position 1 is

$$P(\theta_l|x^{(1)}) = \frac{P(\theta_l)P_1(x^{(1)}|\theta_l)}{P_1(x^{(1)})} \quad (3)$$

Using the RHS of equation (3) as the prior for calculating the posterior in position 2 gives

$$\frac{\frac{P(\theta_l)P_1(x^{(1)}|\theta_l)}{P_1(x^{(1)})}P_2(x^{(2)}|\theta_l)}{P_2(x^{(2)})} = \frac{P(\theta_l)P_1(x^{(1)}|\theta_l)P_2(x^{(2)}|\theta_l)}{P_1(x^{(1)})P_2(x^{(2)})} \quad (4)$$

Using the RHS of equation (4) as the prior for calculating the posterior in position 3 gives

$$\frac{\frac{P(\theta_l)P_1(x^{(1)}|\theta_l)P_2(x^{(2)}|\theta_l)}{P_1(x^{(1)})P_2(x^{(2)})}P_3(x^{(3)}|\theta_l)}{P_3(x^{(3)})} = \frac{P(\theta_l)P_1(x^{(1)}|\theta_l)P_2(x^{(2)}|\theta_l)P_3(x^{(3)}|\theta_l)}{P_1(x^{(1)})P_2(x^{(2)})P_3(x^{(3)})} \quad (5)$$

Continuing on in this fashion through the p^{th} position yields

$$\frac{P(\theta_l)P_1(x^{(1)}|\theta_l)P_2(x^{(2)}|\theta_l) \cdots P_p(x^{(p)}|\theta_l)}{P_1(x^{(1)})P_2(x^{(2)}) \cdots P_p(x^{(p)})} = P(\theta_l|x^{(1)}, \dots, x^{(p)}) \quad (6)$$

which is the desired result. ■

This calculation can run until the the end of the barcode is reached, or it may be terminated early via some kind of stopping rule. The benefit of implementing a stopping rule will be discussed in Section 3.1. Upon reaching the end of the barcode or the stopping rule, the new barcode is then classified as belonging to the species with the highest posterior probability, or it is determined that the new barcode does not belong to any of the species in the reference data set.

3.1 Stopping Rules

If a new barcode can be accurately assigned to one of the species in the reference data set with a shorter barcode, calculations can be sped up and costs decreased. This question could be entertained by the proposed method by implementing a simple stopping rule. Once the posterior probability for a particular species gets sufficiently close to one, the calculations are stopped, and the barcode is assigned to the species yielding that posterior probability. By noting the position upon which the calculation stopped, one might begin to assess necessary barcode lengths for proper classification.

Another important feature that could be “built in” to the proposed method of classification is the process of identifying new or rare species. If a barcode for a species not contained in the dataset is classified, it is reasonable to think that the posterior probabilities of the species in the reference data set should all be around $1/s$. Convergence of the posterior probabilities to $1/s$ could then serve as a basis for attempting to discover new or rare species. If all of the posterior probabilities get sufficiently close to $1/s$, calculation stops and the barcode is classified as not belonging to any of the species in the training data set. More on how this could be achieved with the proposed method will be discussed in section 3.5.

3.2 Adjusting the Conditional Probabilities

By constructing the conditional probabilities as in Section 3, it is clear that if the nucleotide in position j of the new barcode does not occur in any of the observations for a species on position j , the resulting posterior probability of the barcode belonging to that species will be zero, and all subsequent posterior probabilities for that species will be zero. In Section 1.2, it was observed that, while barcodes for the same species will be very similar, there may be some small amount of variation making this strict specification of the conditional probabilities too rigid. A single position containing a value not observed in the dataset for a particular species will provide a penalty so severe that the posterior will never recover. To avoid this potential problem, we recommend adjusting the conditional probabilities slightly by assigning to all of the conditional probabilities that would be zero some small bit of mass, say δ , where $0 < \delta \ll 1$. The amount of mass distributed to these zero valued conditional probabilities must be taken out of the non-zero valued conditional probabilities so that the probabilities still sum to one. While the calculation of the posterior probability would certainly reflect that the new barcode contained a value not observed by any species in the reference data set, it would, nevertheless, allow the posterior to “recover” if subsequent matches are made, or drive the calculation in equation (1) to zero if subsequent matches are not made.

3.3 Prior Specification

How the prior probabilities, $P(\theta_l)$, are specified is typically a subjective issue, but in the case of DNA barcoding, one can hope that the data will eventually dominate reasonable priors. Due to the sequential calculation illustrated in Section 3, the specified prior probabilities

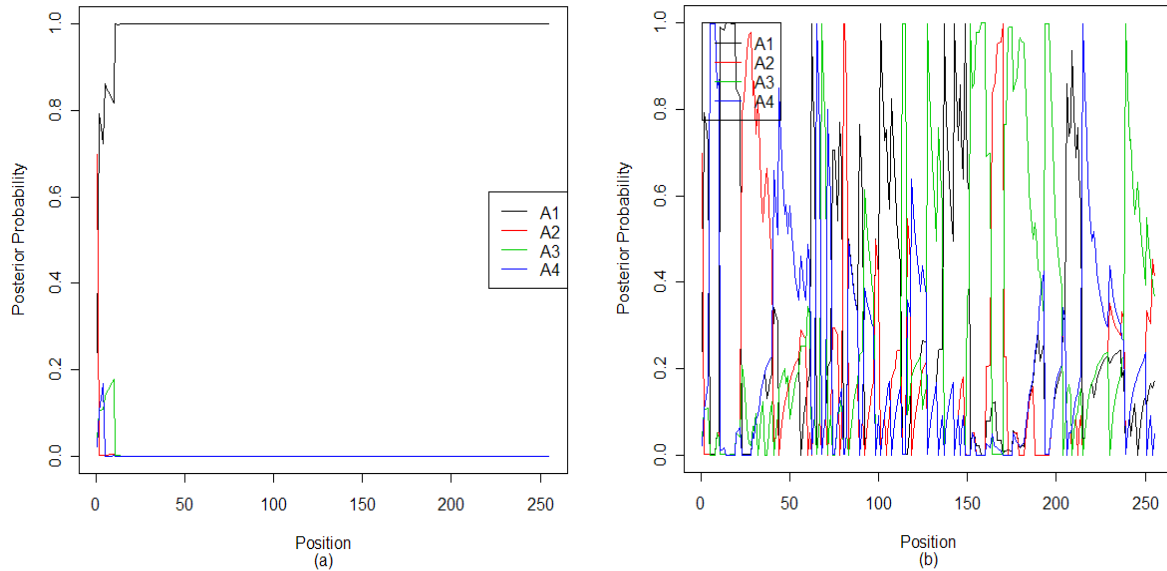


Figure 1: Plotted posterior calculations for each of the four species A1, A2, A3, and A4 at each of 255 positions using randomly generated non-informative Dirichlet priors of $P(A1) = 0.287$, $P(A2) = 0.629$, $P(A3) = 0.058$, and $P(A4) = 0.025$ (a) Typical classification of a new barcode that belongs to the reference data set, (b) Typical classification of a new barcode that does not belong to the reference data set. *Priors do not sum to one because of rounding.*

are updated by the data at each position. Thus, it is expected that the resulting posterior probabilities calculated by equation (1) will be somewhat robust to the choice of priors. Sensitivity analyses that compare non-informative Dirichlet priors with priors based on the proportion of each species in the reference data set, as well as equal priors set to $1/s$ for each species yield nearly identical results. See Table 3.

3.4 Missing Data

The methods used to retrieve DNA barcodes, while very good, are not infallible. Occasionally, the process will not be able to identify the base at a particular location, or more frequently, the process of aligning the barcodes so that their positions match-up yields positions for which no base has been observed. Each of these situations leads to missing data at various positions along the barcode.

Whether the data is missing due to software induced alignments or to the presence of an ambiguous base during extraction, it will have an impact on whether the method can properly classify the barcode. Note that data missing to these causes can show up in the reference data sets used to create the conditional probabilities discussed in Section 3, as

well as in the barcodes to be classified. If a base is missing in a position for a barcode in the reference data set, it might be possible to impute it. This will be discussed presently. However, a missing base from a barcode to be classified will have no grounds upon which it might be imputed. In this case, it is advisable to simply skip that position and use only the positions for which nucleotide information is available.

If a base is missing at a particular position from a barcode in the reference data set, it may be possible to use the barcodes from other organisms within that species to tell us something about what base should have been observed. The missing base may be imputed by either assigning to it the value of the most frequently occurring base in that position for the particular species, or it may be randomly assigned one of the four base values with probabilities equal to their observed proportions within that species at that position. Both methods yield nearly identical imputations and misclassification rates (results not shown). If all of the bases are missing in a position for a particular species, there is no information from which to impute the missing values. The conditional probabilities at these positions are missing and should not contribute to the calculation of the posterior probabilities. Therefore, the conditional probabilities at these positions should simply be set to $1/4$ in the calculation of equation (1).

3.5 Species Discovery and the Proposed Method

Equation (1) can classify an unknown barcode to one of the species in the reference data set, but it does not identify new barcodes that do not belong to the reference data set. Consider the following simple adjustment to the posterior probability of species l , $P(\theta_l|\cdot^{(j)})$, when the base at position j in the barcode to be classified does not match any of the bases in position j of the barcodes in the reference data set:

$$P(\theta_l|\cdot^{(j)}) = P(\theta_l|\cdot^{(j-1)}) + (1/s - P(\theta_l|\cdot^{(j-1)})) \cdot \epsilon, \quad (7)$$

for $l = 1, \dots, s$, where s is the number of species in the reference data set, ϵ is a value between 0 and 1 that reflects the rate of convergence of the posterior probability to $1/s$, and $P(\theta_l|\cdot^{(j-1)})$ is the prior probability of species l for position j . Larger values of ϵ would cause $P(\theta_l|\cdot^{(j)})$ to converge to $1/s$ at a faster rate while smaller values of ϵ would cause $P(\theta_l|\cdot^{(j)})$ to converge to $1/s$ at a slower rate. The convergence of $P(\theta_l|\cdot^{(j)})$ to $1/s$ is easily seen by rewriting the right hand side as

$$P(\theta_l|\cdot^{(j-1)}) + (1/s - P(\theta_l|\cdot^{(j-1)})) \cdot \epsilon = P(\theta_l|\cdot^{(j-1)}) \cdot (1 - \epsilon) + (1/s) \cdot \epsilon \quad (8)$$

One way to view equation (8) is as a weighted sum of the prior probabilities and $1/s$. If a large value of ϵ is chosen, the contribution of the prior probabilities to the posterior probabilities gets down weighted while the contribution of $1/s$ to the posterior probabilities is increased. The value of ϵ used here can be thought of as something similar to the smoothing constant used in single exponential smoothing.

Posterior probabilities will, therefore, tend to $1/s$ when the barcode to be classified does not match any of the bases in the reference data set. If a new barcode to be classified does

Data Set	s	k	R	T	M_D	M_P	M_E	M_C	\bar{p}
Bat	96	659	756	84	0	0	0	0.012	148.078 (153.368)
Bird1	150	690	1461	162	0.0019	0.0019	0.0019	NA	138.238 (123.549)
Bird2	289	255	2330	259	0.0317	0.0289	0.0280	0.142	190.239 (61.389)
Butterfly	205	255	3839	427	0.0062	0.0065	0.0065	0.0962	208.309 (49.798)
Fish	112	255	678	76	0.00776	0.0076	0.0060	0.014	166.496 (63.189)

Table 3: Misclassification Rates for non-informative Dirichlet, Proportional, and Equal Priors for Proportional Allocation Imputation. R and T are the number of barcodes in the reference and test data sets, respectively. M_D , M_P , and M_E represent the misclassification rates for the non-informative Dirichlet, proportional, and equal priors, respectively and M_C represents the misclassification rates for the current method. The column \bar{p} gives the average number of positions (rounded to the nearest whole number) required by the proposed method to classify the barcodes using non-informative Dirichlet priors along with the standard deviation. In each case $\delta = 0.0001$.

not belong to any of the species in the reference data set, but has bases that match the bases of the reference data set, the species with the highest posterior probability will change frequently as the calculation in equation (1) proceeds. A plot of the posteriors for all species versus the number of positions used would then prove useful in detecting new or rare species. Figure 1 shows the plotted posterior probabilities for (a) a typical plot in which the new barcode was properly classified and (b) a barcode that does not belong to the reference data set. The plot in (a) would have terminated around the 35th position according to the early stopping rule, but the stopping rule was removed so that the posterior probabilities would be plotted on the same scale as those in (b) which did not trigger the stopping rule and therefore cover all of the positions on the barcode.

3.6 Algorithm of the Proposed Method

The following algorithm outlines the process of the proposed method of classification:

Based on a reference data set of barcodes R and a new barcode T :

1. Impute the missing data in R as discussed in Section 3.4.
2. Using R , compute the conditional probability of the bases A, T, C, and G for every species at every position.
3. Adjust the conditional probabilities above by assigning δ to all zero valued conditional probabilities while adjusting the nonzero conditional probabilities so that they will still sum to 1.
4. Select prior probabilities $P(\theta_l)$ for each species $l = 1, \dots, s$.

5. If the base in position j of T is missing, skip to position $j + 1$. Otherwise, continue to the next step.
6. If the base in position j of T does not match any of the bases in position j of R , use equation (7) to calculate the posterior probabilities for each species. Otherwise, use equation (1) to calculate the posterior probabilities for each species.
7. Repeat (5) and (6) until any of the following occur
 - (a) If $P(\theta_l|\cdot^{(j)}) = 1$, for any $l = 1, \dots, s$, stop and classify barcode to species θ_l .
 - (b) If $P(\theta_l|\cdot^{(j)}) = 1/s$ for all $l = 1, \dots, s$, stop and conclude the new barcode does not belong to any species in R .
 - (c) The end of the barcode is reached. At this point, posterior probabilities at each position should be plotted and examined. If the plot shows the species with the highest posterior probability frequently changing like the “noisy” plot in Figure 1 (b), conclude the new barcode does not belong to any species in R . If, however, the plot clearly favors one of the species over the rest like the plot in Figure 1 (a), classify the barcode as belonging to the species θ_l with the highest computed posterior probability.

4 Results

To explore the effectiveness of the proposed method, five data sets containing barcode data were extracted from BOLD and analyzed. 10-fold crossvalidated misclassification rates were computed for various choices of priors by splitting each data set into a reference data set R , consisting of around 90% of the observations, and a test data set T consisting of the remaining 10%. The splitting of each data set consisted of randomly selecting the desired percentage of observations to make up the test data set. After this randomization, care was taken to ensure that each species had at least one representative in the reference data set. The barcodes in the test data set were then classified using the proposed method with the missing data being imputed via the proportional allocation approach and arbitrarily choosing $\delta = 0.0001$.

Table 3 provides misclassification rates for the proposed method when non-informative Dirichlet priors, M_D , priors proportional to the number of organisms per species in the reference data set, M_P , and equal priors of $1/s$ for each species, M_E , were used. M_C are the misclassification rates for the current method (Gusev, Kentros, Lindsay, Mandaoui, and Pasaniuc 2007). It is interesting to note that the misclassification rates for the proposed method seem to be somewhat robust to the choice of prior probabilities used, supporting the claim that the barcode data should eventually overcome reasonable priors to provide nearly identical misclassification rates. It is also interesting to note the average number of positions required for the classification in the last column labeled \bar{p} . For the Bat and Bird1 data sets, with the largest number of positions, 659 and 690, respectively, the entire

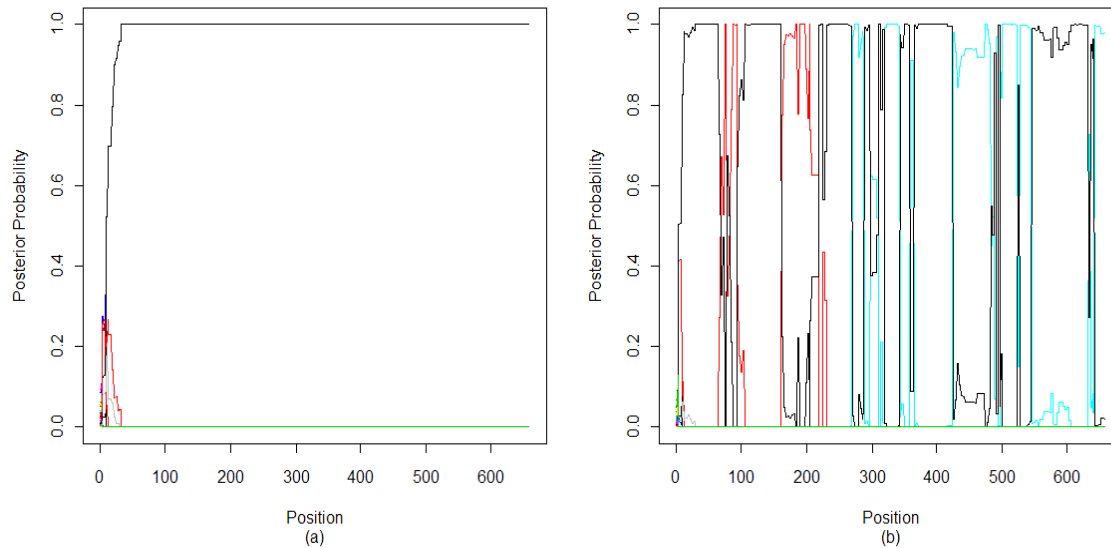


Figure 2: Plotted posterior calculations for the bat data set where (a) all of the 96 species were used in the reference data set with one barcode from the species *Uroderma bilobatum* held out of the reference data set and classified; (b) the species *Uroderma bilobatum* was completely removed from the reference data set and a barcode from that species was classified.

barcode was almost never necessary for classification, and only 148 and 138 positions were used on average, respectively. In each data set, we observed that proper classification can be performed within about 200 positions. The Bird2 data set was a large data set with several barcodes that had many missing positions which increased misclassification for the proposed method to around 3% and for the current method to around 14%. The Butterfly data set was unique in that the intra-species variability was much smaller (about half) compared to the other animals. For current methods, this narrow genetic “gap” made classification for these species less distinguishable and more difficult and misclassification rates jumped to around 10%, whereas the proposed method maintained low misclassification rates of less than 1%.

To examine the ability of the proposed method to indicate that a new barcode does not belong to any of the species in the reference data set, we randomly selected the species *Uroderma bilobatum* and completely removed it from the bat reference data set. We then attempted to classify a barcode from that species using non-informative Dirichlet prior probabilities with δ arbitrarily set to 0.0001, and a small value of ϵ chosen to be 0.2. Figure 2 (b) is a plot of the posterior probabilities of each species at each position. The species with the highest posterior probability fluctuates between several species and none of them get close enough to unity to trigger the stopping rule discussed in Section 3.1. This is a clear indication that the new barcode represents a new species not contained in the reference data

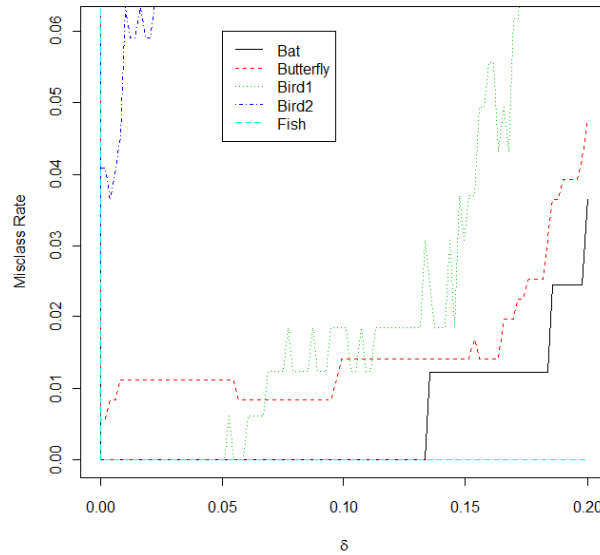


Figure 3: Plotted misclassification rates for the five data sets for various δ values. This plot shows that misclassification rates increase if δ is too close to 0 or if it is larger than 0.01.

set. Compare this plot to Figure 2 (a) where the species *Uroderma bilobatum* was in the reference data set and the new barcode belonged to that species. This plot shows the posterior probabilities clearly favoring a single species from around the 20th position on. Using the stopping rule, this classification would have terminated around the 60th position.

5 Summary

It has been shown here that the proposed method performs reasonably well compared to the current method in terms of classification. In addition, this method provides a theoretically sound and easily interpretable probability for each classification, while avoiding genetic/evolutionary model based assumptions. It also begins to address the issues of necessary barcode length and species discovery.

A matter of further investigation is the effect a particular δ value may have on misclassification rates. Figure 3 shows how sensitive misclassification rates of the proposed method are for various δ values. It is clear that if the δ value is too close to zero, the misclassification rates increase. Likewise, misclassification rates appear to increase if the δ value gets too large. For these data sets, it appears that supplying the proposed method with a δ value in the interval (0,0.01) yields lower misclassification rates. These results indicate that the misclassification rates for the proposed method in Table 3 depend on data-set-specific δ values. One biological interpretation for δ is the probability of observing a mutation at any of

the positions within the COI region and an optimal choice for δ may be closely related to estimates of the mutation rate in this region.

References

- Altschul, S. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- DeSalle, R. (2006). Species discovery versus species identification in dna barcoding efforts: response to rubinoff. *Cons. Biol.* 20, 1545–1547.
- Ferguson, J. (2002). On the use of genetic divergence for identifying species. *Biol. J. Linn. Soc.* 75, C509–C516.
- Frézal, L. and R. Leblois (2008). Four years of dna barcoding: Current advances and prospects. *Infect. Genet. Evol.* 8(5), 727–736.
- Gusev, A., S. Kentros, J. Lindsay, I. Mandaoiu, and B. Pasaniuc (2007). A comparison of algorithms for species identification based on dna barcodes (invited talk). In *2nd International Barcode of Life Conference*, Academia Sinica, Taipei, Taiwan.
- Hajibabaei, M., J. DeWaard, N. Ivanova, S. Ratnasingham, R. Dooh, S. Kirk, P. Mackie, and P. Hebert (2005). Critical factors for assembling a high volume of dna barcodes. *Phil. Trans. R. Soc. B* 360, 1959–1967.
- Hammond, P. (1992). *Global biodiversity: status of the earth's living resources*. London: Chapman & Hall.
- Hebert, P., S. Ratnasingham, and J. deWaard (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings Biological Sciences* 270, S96–S99.
- Johns, G. and J. Avise (1998). A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. *Mol. Biol. Evol.* 15, 1481–1490.
- Kelly, R., I. Sarkar, D. Eernisse, and R. DeSalle (2006). Dna barcoding using chitons (genus mopalina). *Mol. Ecol. Notes* 7, 177–183.
- Kimura, M. (1980). The neighbor-joining method: A new method for reconstruction phylogenetic trees. *J. Mol. Evol.* 16, 111–120.
- Koski, L. and G. Goulding (2001). The closest blast hit is often not the nearest neighbor. *J. Mol. Ecol.* 52, 540–542.
- Meyer, C. and G. Paulay (2005). Dna barcoding: Error rates based on comprehensive sampling. *Plos. Biol.* 3, 2229–2238.
- Ratnasingham, R. and P. Hebert (2007). Bold: The barcode of life data system. *Mol. Ecol. Notes* 7, 355–364.
- Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstruction phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Stoeckle, M. (2003). Taxonomy, dna, and the bar code of life. *Bioscience* 53(9), 2–3.

Will, K. and D. Rubinoff (2004). Myth of the molecule: Dna barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20, 47–55.