Michael P. Anderson* and Suzanne R. Dubnicka

A sequential naïve Bayes classifier for DNA barcodes

Abstract: DNA barcodes are short strands of 255–700 nucleotide bases taken from the cytochrome c oxidase subunit 1 (COI) region of the mitochondrial DNA. It has been proposed that these barcodes may be used as a method of differentiating between biological species. Current methods of species classification utilize distance measures that are heavily dependent on both evolutionary model assumptions as well as a clearly defined "gap" between intra- and interspecies variation. Such distance measures fail to measure classification uncertainty or to indicate how much of the barcode is necessary for classification. We propose a sequential naïve Bayes classifier for species classification to address these limitations. The proposed method is shown to provide accurate species-level classification on real and simulated data. The method proposed here quantifies the uncertainty of each classification and addresses how much of the barcode is necessary.

Keywords: Naïve Bayes classifier; DNA barcoding; phylogenetic analysis; sequential analysis; species classification; species discovery.

DOI 10.1515/sagmb-2013-0025

1 Introduction

Taxonomists face great challenges in the classification and discovery of congeneric, or closely related, species. To determine an organism's species, taxonomy relies upon inspection of an organism's observable morphologic features, such as shapes, sizes, markings, and behaviors.

Sole reliance upon morphologic features to determine species proves challenging for several reasons. First, physical characteristics between two congeneric species may be so similar that they are mistakenly identified as the same species. On the other hand, physical characteristics between two organisms of the same species, such as males and females, may appear quite different, resulting in the classification of two separate species. Further difficulties arise when morphologic features for an organism develop over time, making classification possible only during specific periods of the organism's life cycle.

Second, because morphologic features can only be compared to what has previously been observed, the process of discovering new species can be slow. According to Hebert et al. (2003b), an individual taxonomist can rarely identify more than 1000–1500 different species. Extensive collaboration is then required when observed features do not match those of any species known to a taxonomist before it can be decided a new species has been discovered. The rate of this process can be seen 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 discovered roughly 1.7 million unique species (Stoeckle, 2003).

Lastly, it is sometimes necessary to make species classifications from organism fragments which 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 some birds are attracted to, and collide with, aircraft. Collisions of this type often necessitate classification based on organism fragments (Dove, 2000;

***Corresponding author: Michael P. Anderson,** Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center Oklahoma City, OK, USA, e-mail: michael-anderson@ouhsc.edu
Suzanne R. Dubnicka: Department of Statistics, Kansas State University, Manhattan, KS, USA

Marra et al., 2009), leaving questions about the bird's species difficult to answer using traditional morphologic methods. Also, museums of natural history often contain repositories of unidentified 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 new species, motivate research for a more precise and efficient discrimination among species that can complement the challenges of classification based solely on morphological features.

1.1 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) region of mitochondrial DNA, is an effective marker for identifying species in the animal kingdom (Hebert et al., 2003b). This barcode contains a sequence of the nucleotide bases adenine (A), thymine (T), cytosine (C), and guanine (G) and typically ranges in length from 255 to around 700 nucleotide bases.

Using DNA barcodes to discriminate between species, it is hoped that the previously mentioned challenges in classifying known species, and discovering new ones, can be addressed. To be sure, barcodes can be retrieved from a very small amount of an organism's tissue (1–3 mm^3) and at any stage of life. Obtaining these barcodes is a relatively quick and inexpensive procedure costing $3–5 per barcode (Hajibabaei et al., 2005a). Thus, organism fragments and development or change of morphologic 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 <1% divergence (Johns and Avice, 1998). Hebert et al. (2003b) found that among congeneric species, inter-species barcode divergence averaged 6.8%, with 99.98% having sequence divergence >3%. Barcode divergence was higher 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 dipsarity between a novel barcode to be assigned to a species and a set of barcodes that serve as a reference data set.

2 Current methods of Barcode classification

2.1 Distance and similarity measures

Initial attempts to extract information on the relatedness of DNA sequences have resulted in distance measures based on the pairwise differences between two DNA barcodes. In keeping with the theory that barcodes for the same species should be similar, a direct measure of the proportion of dissimilarities between two barcodes known as the p-distance is sometimes used (Hebert et al., 2003a). A more sophisticated distance measure is Kimura's Two Parameter model (K2P) which relies upon the type of difference observed at homologous positions for two nucleotide sequences. If the nucleotides are different but both are purines (A or G) or pyrimidines (C or T), the difference is called a transition while all other discrepencies are called transversions. The genetic distance between sequences *i* and *j* is computed by

$$d(i,j)=-(1/2)ln[(1-2P-Q)\sqrt{1-2Q}] \tag{1}$$

where *P* and *Q* are the proportion of transitions and transversions, respectively (Kimura, 1980).

For both of the above measures, large distances indicate the two barcodes belong to different species while smaller distances favor the possibility that the two barcodes belong to the same species. Phylogenetic trees can be reconstructed using these distance measures together with the neighbor-joining method (Saitou and Nei (1987) amended by Studier and Keppler (1988)). This is an agglomerative clustering algorithm that starts with the branches of the tree equal to the number of barcodes under investigation and joins the most similar barcode pairs at nodes until the branches have been reduced to a single node (Gascuel and Steel, 2006). The resulting tree is a diagram reflecting the genetic relatedness of all the barcodes based entirely on the distance measures provided.

2.1.1 Basic Local Alignment Search Tool (BLAST)

DNA barcodes are often stored in large databases, necesitating attention to efficient database searching to classify new barcodes. Altschul et al. (1990) proposes a basic local alignment search tool (BLAST) to return the DNA barcodes from a database that are the most similar to a query barcode. The maximal segment pair (MSP) score is computed for any two sequences using only highly conserved regions of the DNA. Typically, the MSP is computed by giving a +5 score to identical base-pairs and a −4 score to mismatches. For two aligned contiguous segments of base-pairs of equal length in a highly conserved region, the similarity score is the sum of the similarity values for each position compared in the segment. The MSP is the highest scoring pair of identical length segments from 2 sequences: one of which is the query sequence and the other is a sequence from the database.

Results from Karlin and Altschul (1990) allow for estimation of the highest MSP score, say T, for which chance similarities are likely. The search can then minimize the time that it spends on segments that are not likely to exceed this score in terms of their similarity to the query sequence. In practice, a "word," or w-mer, is a segment with fixed length w, and the BLAST search can find segments that contain words with MSP scores of at least T. Focus is then limited to these segments in that they represent significant biological relationships as opposed to chance similarities.

Gains in computational time are enhanced with a filter that removes noisy words from the query sequence. This is done by computing, *a priori*, frequencies of all words, for a fixed word length w, from the DNA database. The words that occur much more frequently than expected by chance are stored and used to filter out the remaining "noisy" words produced by the query sequence. A BLAST search performed with w=12 can scan about 2×10^6 bases/sec (Altschul et al., 1990).

2.2 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 et al. (2006) and Frézal and Leblois (2008). Classification of a novel barcode using the BOLD system proceeds as follows. First, a BLAST search of the BOLD database is implemented to retrieve barcodes from a 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 database. Distance measures, such as K2P, which is the default, are then 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 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 important questions unanswered and has severe limitations. First, it is prone to high rates of false matches in that it will classify the new barcode to its closest neighbor in the tree 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 to the species to which it was classified cannot be measured. Rather, percentages of similarity are reported, which lack solid probabilistic interpretation at the species level and have been shown by Ferguson (2002) to be somewhat unreliable. Third, the distance measures discussed earlier 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 genetic gap, upon which the efficacy of distance measures is predicated, may not be well-separated when comprehensive data sets are considered. They show that using a reference data set containing just a few observations per species (1–2 individuals) severely underestimates intra-species variation, and further argue that there may be much more overlap between intra- and inter-species variation than was previously supposed. The accuracy and overall relevance of these classification methods, which depend on this clearly defined gap, then come into question. Finally, current methods do 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 database, but little justification as to this particular length is provided. Sequences of this length from fresh DNA are easily obtained, but often for archival DNA more than a decade old, contiguous sequences of more than 300–400 base positions are rare (Hajibabaei et al., 2005a).

As an alternative, we propose a sequential naïve Bayes (NB) classification method that assigns theoretically sound probabilities to all positive classifications (Section 3). This method relies less heavily upon the genetic gap than current methods and does not make any genetic/evolutionary model assumptions (Section 4.1). It also provides information as to how much of the barcode is necessary for proper classification (Section 5).

2.3 Naïve Bayes classifier

As the name implies, this classifier utilizes Bayes' rule to classify an observation as belonging to one of s groups. "Naïve" in this sense means that this classifier will use strong independence assumptions that are perhaps unwarranted. Specifically, the NB classifier assumes conditional independence among all of the predictor variables of an observation. In the case of DNA barcoding, this amounts to assuming that the nucleotide positions of a barcode are conditionally independent given species. More succinctly, if the group or class is denoted by S, and $x^{(1)}, \ldots, x^{(p)}$ are p predictor variables, then the NB classifier is

$$class(x^{(1)},\ldots,x^{(p)}) = \arg\max_{S} P(S_s)\prod_{j=1}^{p} P(x^{(j)}\,|\,S_s) \tag{2}$$

It can be shown that this classifier is proportional to the highest posterior probability when the predictor variables are conditionally independent. One attractive aspect of the NB classifier is that the independence assumptions allow for estimating likelihoods of univariate distributions for each predictor variable, making it well-suited for high dimensional data. Ignoring the dependent structure of the predictor variables can lead to poorly estimated group probabilities but still result in the correct classification as long as the correct group is the most probable. Zhang (2004) discusses the implications of these independence assumptions and the NB classifier's effective performance even when such assumptions are questionable. McCallum and Nigam (1998) provide a nice overview of naïve Bayes classifiers, particularly as they apply to text classification.

3 Proposed method of classification

Let $x^{(1)}, \ldots, x^{(p)}$ represent the observed nucleotides in a barcode sequence with p positions, and S_l be the event that the barcode to be classified belongs to species l, where l=1, …, s, and s is the number of species in the reference data set. We define $P(S_l)$ to be the prior probability the barcode belongs to species l, and $P(x^{(j)}|S_l)$ to be the conditional probabilities of the nucleotide values A, T, C, and G at each position j=1, …, p given each species in the reference data set. We estimate these conditional probabilities with 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 3 different species. For the first three positions of species 1, the estimated conditional probabilities are given in Table 2.

With these specifications, posterior probabilities of the barcode belonging to each of the s species are calculated sequentially at each position according to the following equation:

$$P(S_l|x^{(1)},\dots,x^{(j)})=\frac{P(S_l|x^{(1)},\dots,x^{(j-1)})P(x^{(j)}|S_l)}{\sum_{l=1}^{s}P(S_l|x^{(1)},\dots,x^{(j-1)})P(x^{(j)}|S_l)} \quad (3)$$

We point out that $P(S_l|x^{(1)}, \dots, x^{(j-1)})$ is the estimated posterior probability that the barcode belongs to species l after observing the nucleotides at position 1, …, j–1 and serves as the prior probability in the calculation of $P(S_l|x^{(1)}, \dots, x^{(j)})$, the posterior probability that the barcode belongs to species l after observing the nucleotides at positions 1, …, j. For j=1, the value $P(S_l|x^{(0)})$ is equal to the specified prior $P(S_l)$. The goal of equation (3) is to update the original prior probability at each nucleotide position and ultimately estimate $P(S_l|x^{(1)}, \dots, x^{(p)})$. We evaluate the performance of this classifier when the functional form of $P(S_l)$ is: 1. randomly generated from a *Dirichlet*(1, …, 1), 2. generated from a *Discrete Uniform* distribution, and 3. proportional to observed data. We now state a proposition and its corollary regarding the proposed method and the resulting calculated posterior probabilities.

Proposition 1 *Let $x^{(1)}, \dots, x^{(p)}$ be p independent observations that arise in sequence. Let $P(S_l)$ represent the prior probability that the sequence of observations belongs to group l, and suppose that the sequence of observations are conditionally independent, given group l. Suppose further that the conditional probabilities $P(x^{(1)}|S_l), \dots, P(x^{(p)}|S_l)$ are known.*

Then, by using the posterior probability from position j, $P(S_l|x^{(1)}, \dots, x^{(j)})$, as the prior probability in equation (3) for computing the posterior at position j+1, $P(S_l|x^{(1)}, \dots, x^{(j+1)})$, sequentially for j=1, …, p, results in computing $P(S_l|x^{(1)}, \dots, x^{(p)})$.

Corollary 1 *Let $x^{(1)}, \dots, x^{(p)}$ be p independent nucleotides that, in sequence, constitute a DNA barcode. Let $P(S_l)$ represent the prior probability that the barcode belongs to species l. Suppose that the nucleotides are conditionally independent, given species l and that the conditional probabilities, $P(x^{(1)}|S_l), \dots, P(x^{(p)}|S_l)$, are known. Suppose further that the barcodes within each species are identical, and the species in the reference data set, R, represent all possible species and have unique barcodes.*

Then the posterior probability of belonging to species l, $P(S_l|x^{(1)}, \dots, x^{(p)})$, as estimated by equation (3) will be monotone increasing if and only if the new barcode T to be classified belongs to species l.

Proposition 1 states that under the assumptions of independence among the nucleotide positions, updating the initial prior probability sequentially along the entire sequence results in an unbiased estimate of the true probability of belonging to group l. Corollary 1 states that under somewhat strict uniqueness assumptions, the estimated posterior probability for the correct group is strictly non-decreasing while the posterior probabilities for the incorrect groups cannot be non-decreasing. The proofs of Proposition 1 and its corollary are given in the Appendix.

4 Methods and results

To explore the effectiveness of the proposed method, 12-fold cross-validation was performed on simulated barcode data sets, and 10-fold cross-validation was performed on real barcode data sets. Four simulated data sets were created with 2, 4, 6, and 8% intra-species variablitiy and 7–10% inter-species variability. This was done to assess the accuracy of the proposed method when the genetic gap is and also when it is not well-defined. Five real data sets were extracted from BOLD and contain barcodes for bats, birds, butterflies, and

Table 1 Truncated Barcode data.

Species	Truncated Barcode
1	CCGGCATAGTAGGCACTG
1	CCGGCATAGTAGGCACTG
1	CCGGCATAGTTGGCACTG
1	CTGGCATAGTAGGTACTG
2	CCGGCATAGTAGGAACAG
2	CTGGCATAGTAGGAACAG
2	CCGGCATAGTAGGAACAG
3	CCGGAATAGTAGGTACCG
3	CCGGAATAGTAGGTACCG
3	CCGGAATAGTAGGTACCG

Barcodes are short DNA sequences made up of the nucleotide bases adenine (A), thymine (T), cytosine (C), and guanine (G). The sequences are taken from the COI region of the mitochondrial DNA and may range in length from 255 positions, to 700 positions. The first 18 positions of the barcodes retrieved from 10 bats belonging to 3 different species are shown here.

Table 2 From the four bats belonging to species 1 in Table 1, the conditional probabilities of observing any of the four nucleotides, given their species affiliation, for the first 3 positions of species 1 (S_1) are given here.

Position 1	Position 2	Position 3
$P(A\|S_1)=0/4$	$P(A\|S_1)=0/4$	$P(A\|S_1)=0/4$
$P(T\|S_1)=0/4$	$P(T\|S_1)=1/4$	$P(T\|S_1)=0/4$
$P(C\|S_1)=4/4$	$P(C\|S_1)=3/4$	$P(C\|S_1)=0/4$
$P(G\|S_1)=0/4$	$P(G\|S_1)=0/4$	$P(G\|S_1)=4/4$

fish. These data sets are publicly available at the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS, 2007).

10-fold cross-validated misclassification rates were computed using various methods for selecting prior probabilities. Prior probabilities were randomly generated from a Dirichlet distribution (P_D), chosen to be proportional to species representation in the data (P_P), or set equal to one another (P_E). Each data set was randomly split into a reference data set and a test data set where care was taken to ensure that the reference data contained at least one representative from each species.

We adjusted the conditional probabilities used in equation (3) by assigning the zero-valued probabilities a small amount of mass, δ, where $0<\delta\ll 1$. In the context of DNA barcoding this quantity can be thought of as the mutation rate of nucleotides in the COI region. For the real data examples, we set this quantity equal to 9.7×10^{-8} which is based on Denver et al. (2000) who propose it as an estimate of the mutation rate of the *COI* region for *Caenorhabiditis elegans*. This adjustment permits rare mutations in the barcode to be observed, without setting to zero the posterior probability of its true species affilitation.

4.1 Simulated data

A simulated barcode data set was created by generating a single DNA sequence with 700 nucleotide positions, p, with nucleotide proportions equal to those observed in a real barcode data set. From this "seed" barcode, 12 sequences were generated representing 12 unique species such that they had exactly 8% inter-species varaibility. From each of these 12 barcodes, four sequences were generated representing barcodes from 4 distinct organisms within the same species for a total of 48 barcodes. The four sequences within each species were generated such that they had exactly 2% nucleotide variability in accordance with Avise (2000). Setting the within-species variability in this fashion altered the among-species variability slightly for each pair of species which ranged between 7 and 10%. We repeated this process to obtain three additional data sets having within-species variability of 4, 6, and 8%, while maintaining among-species variability of 7–10%.

The results, given in Table 3, illustrate the misclassification rates of the proposed method are robust to the choice of priors. To explore how much of the barcode is necessary for accurate classification, a stopping rule was triggered for the proposed method when the posterior probability for any species reached unity. The average number of positions examined for classification, $\bar{p}$, using equal priors ranges from 60 to 140, demonstrating how narrowing the genetic gap requires the proposed method to use more nucleotide positions for the classification. This also demonstrates how the current method (*K2P*) is more sensitive to a clearly defined gap than the proposed method. Within-species variability of 8% causes the current method to misclassify a third of the barcodes, on average, as opposed to the proposed method's average misclassification rate of a fourth. These misclassification rates are based on an average of 700 nucleotide positions for the current method, and an average of 140 nucleotide positions for the proposed method.

4.2 Real data

The real data sets analyzed in this study come from barcodes that have been submitted to the Barcode of Life (BOLD) repository. The two bird data sets, Bird1 and Bird2, consisted of barcodes from 1623 and 2563 organisms, respectively, accounting for 150 and 289 unique species, respectively. Likewise, the Bat, Butterfly, and Fish data sets consisted of barcodes from 840, 2563, and 750 organisms accounting for 96, 205, and 112 unique species. The barcodes ranged in length from 255 nucleotides in the Bird2, Butterfly and Fish data sets, to 659, and 690 nucleotides in the Bat and Bird1 data sets, respectively. Detailed documentation on the Bat, Bird, Butterfly, and Fish data sets along with their BOLD ascension numbers can be found in Clare et al. (2006), Kerr et al. (2007), Hajibabaei et al. (2005b), and Ward et al. (2005), respectively. Links to these data sets in spreadsheet format, as well as links to the references given above are available at http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007/ DIMACS (2007).

Many of the barcodes in the real data sets contained missing data at the beginning or end of the barcode due to sequence alignments. We chose to impute the missing values within a species from the observed proportion of nucleotides of organisms within the species. Another approach was considered where the missing values were imputed with the most frequently occurring nucleotide within the species. These two approaches yielded nearly identical misclassification rates (Supplementary Material Table 5). When every nucleotide for a given position within a species was missing, the conditional probabilities for that species at that position were set to 1/4.

To address how much of the barcode is necessary for proper classification, we implement a stopping rule when the computed posterior probability for any species gets sufficiently close to unity, at which point the calculation is terminated and the classification is made to the species yielding the highest posterior probability. Misclassification rates were assessed with and without the early stopping rule and proved to be nearly identical (Supplementary Material Table 6).

Table 3 Misclassification rates for the four simulated barcode data sets with 2%, 4%, 6%, and 8% intra-species variability.

Data set	*S*	*p*	P_D	P_P	P_E	*K2P*	$\bar{p}$ (*SD*)
2%	12	700	0	0	0	0	60.611 (24.126)
4%	12	700	0	0	0	0	87.521 (39.828)
6%	12	700	0.042	0.042	0.042	0.021	112.05 (93.137)
8%	12	700	0.229	0.25	0.25	0.333	139.063 (83.33)

Each data set has *s* unique species and barcodes length of *p*. P_D, P_P, P_E, are the misclassification rates for the proposed method using Dirichlet generated, data proportional, and equal priors, respectively. K2P is the misclassification rate using Kimura's 2 parameter model. The average number of nucleotide positions used by the proposed method for classification along with the standard deviation are given by $\bar{p}$(*SD*).

Figure 1 illustrates the posterior probabilities computed using the proposed method at each position for a randomly selected bat barcode belonging to the species *Artibeus obscurus* from the bat data set. Starting with randomly generated *Dirichlet* priors, the proposed method of classification estimates the posterior probability of the barcode belonging to each species at each nucleotide position. From the plot we see the species *Artibeus obscurus* is identified as a likely match with posterior probability close to 0.55 around the 70th position with a positive classification being made around the 225th position.

Table 4 displays the misclassification rates for all five data sets and the three choices of priors. The average number of nucleotide positions required to make the classification using equal priors is given in the last column of the table and ranges from 85 for the first bird data set to 170 for the butterfly data set. The misclassification rates for the three choices of priors are very similar across all five data sets. The proposed method yields improved misclassification rates over the current method using K2P distances for the Bird2, Butterfly, and Fish data sets. While the current method provides better misclassification rates for the Bat and Bird1 data sets, using an arbitrarily selected δ value of 1.0×10^{-4} with the proposed method reduces the

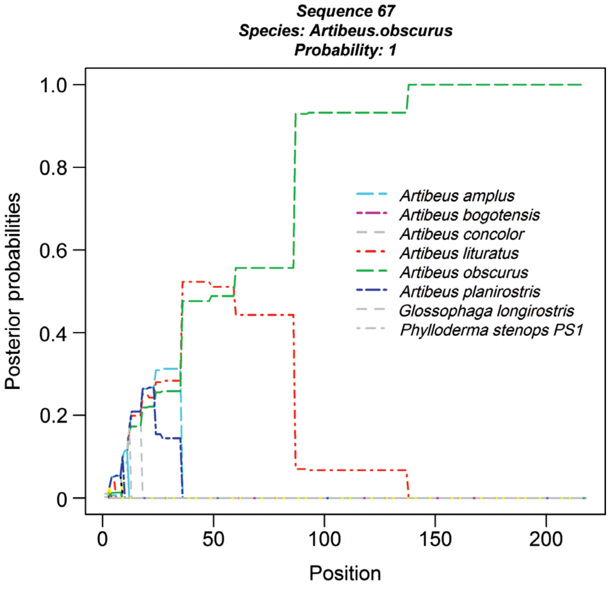

Figure 1 Plotted posterior probabilities for a bat barcode belonging to the species *Artibeus obscurus*. For all 96 bat species, these probabilities are plotted across barcode positions by different line types. Using randomly generated *Dirichlet* priors the proposed method identifies the correct species (*Artibeus obscurus* is the dashed green line) by the 70th position and the stopping rule is triggered at the 225th position. The barcode being considered, species to which it was classified, and estimated probability of species assignment are given in the figure heading.

Table 4 Misclassification rates for the five real barcode data sets each with *n* barcodes.

Data set	*n*	*s*	*p*	P_D	P_P	P_E	*K2P*	$\bar{p}$(*SD*)	95% CI
Bat	840	96	659	0.001	0.001	0.001	0	98.744 (103.122)	(91.65, 105.89)
Bird1	1623	150	690	0.003	0.003	0.003	0.001	84.49 (64.34)	(81.30, 87.68)
Bird2	2564	289	255	0.025	0.021	0.023	0.029	156.376 (63.09)	(153.88, 158.87)
Butterfly	4224	205	255	0.005	0.005	0.005	0.056	169.886 (55.198)	(168.19, 171.58)
Fish	750	112	255	0.008	0.008	0.006	0.015	137.76 (62.176)	(133.22, 142.30)

Each data set has *s* unique species and barcodes length of *p*. P_D, P_P, P_E, are the misclassification rates for the proposed method using *Dirichlet* generated, data proportional, and equal priors, respectively. K2P is the misclassification rate using Kimura's 2 parameter model. The average number of nucleotide positions used by the proposed method for classification along with the standard deviation are given by $\bar{p}$ (*SD*). (95% confidence intervals for the average number of positions required for classification are also given.)

misclassification rates to 0 and 0.002, respectively. This may be an indication that using a species-specific estimate of the mutation rate of the COI region could improve misclassification rates.

4.3 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 with s unique species and a new barcode T with p nucleotide positions:

1. Impute the missing data in R.
2. Using R, compute $P(x^{(j)}|S_l)$ for $j=1, \dots, p$ and $l=1, \dots, s$.
3. Adjust the conditional probabilities above by assigning $\delta=9.7\times10^{-8}$, or a more appropriate species specific adjustment, to all zero valued conditional probabilities while adjusting the nonzero conditional probabilities so that they will still sum to 1.
4. Select prior probabilities $P(S_l)$ 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. Use equation (3) to calculate the posterior probabilities for each species.
7. Repeat (5) and (6) until
   (a) $P(S_l|x^{(1)}, \dots, x^{(j)})=1$, for $j<p$ and any $l=1, \dots, s$, stop and classify barcode to species S_l or,
   (b) The end of the barcode is reached. Classify the barcode as belonging to

$$\arg\max_{S} P(S_l \,|\, x^{(1)},\dots,x^{(p)}) \tag{4}$$

5 Summary

DNA barcodes can be effectively used to assign organisms to respective species based on a reference barcode data set. We propose a sequential naïve Bayes classification method that updates species assignment probabilities at each position of the barcode until an identification is made with high probability. It has been shown here that the proposed method can provide significant gains in classification accuracy compared to the current method. Additionally, this method provides theoretically sound and easily interpretable probabilities for each classification, while avoiding genetic/evolutionary model based assumptions making it less sensitive to divergence from such assumptions. It also begins to address the issue of necessary barcode length. For example, on average the Bird1 data set required around 85 nucleotide positions for classification with a standard deviation of 64.34. A Wald-type upper-bound for a 95% confidence interval on the true number of nucleotides required for classification of these birds is about 106, a dramatic reduction from the available 690. The Butterfly data set averaged about 170 nucleotides per classification, the largest of all the data sets, giving Wald-type upper-bound for a 95% confidence interval of 172.

We point out in Section 4.2 the misclassification rates can be sensitive to the choice of δ. Figure 2 illustrates the data sets in this study achieve optimal classification rates at various δ values. While the misclassification rates are not widely different for small values of δ, this implies improved misclassification rates can be achieved by using data specific COI region mutation rates with the proposed method.

To date, current methods focus more on the sensitivity of barcode classification and less, if at all, on the specificity of barcode classification. That is to say, they are more concerned with correctly classifying a barcode that belongs to a species in a reference data set than with correctly indicating when a barcode does not belong to any species in a reference data set. If DNA barcoding is to aid in species discovery as discussed in Section 1, there must be more focus placed on the latter. An important by-product of the method proposed here is the potential it has in addressing the issue of species discovery. For example, a posterior probability is computed for each nucleotide position in the barcode. Supposing the barcode does not belong to any species in the reference data set, the species with the highest posterior probability may change frequently among the

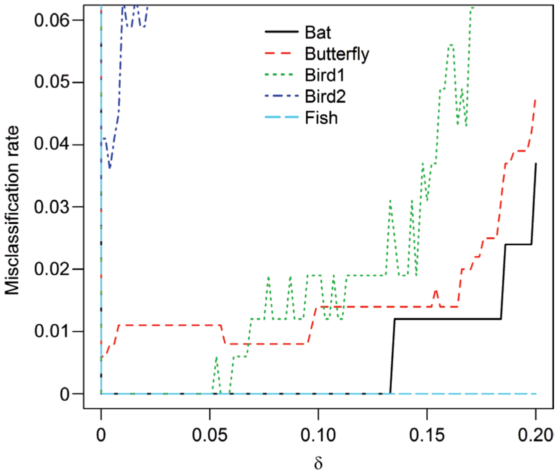

Figure 2 Misclassification rates using the proposed method for the five real data sets at various COI mutation rates, δ. Optimal classification rates for each data set occur at different δ values, suggesting a species specific adjustment may be advantageous.

species within the data set. Plotting the posterior probabilities versus nucleotide position provides a picture of how the final posterior probability was obtained. If the plot shows frequent transitions among species, it may be an indication that the species to which the barcode belongs is not represented in the reference data set. We plan to further investigate how posterior probability plots resulting from the method proposed here may be used for species discovery.

Appendix

Proof of Proposition 1

Proof. First notice that because the observations $x^{(1)} \dots x^{(p)}$ are marginally independent as well as conditionally independent given group l, we have

$$P(S_l|x^{(1)},\dots,x^{(p)})=P(S_l)\prod_{j=1}^{p}P_j(x^{(j)}|S_l)/\prod_{j=1}^{p}P_j(x^{(j)}) \tag{5}$$

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

$$P(S_l|x^{(1)})=P(S_l)P_1(x^{(1)}|S_l)/P_1(x^{(1)}) \tag{6}$$

Using the RHS of equation (6) as the prior for calculating the posterior in position 2 gives

$$\frac{P(S_l)P_1(x^{(1)}|S_l)}{P_1(x^{(1)})}P_2(x^{(2)}|S_l)/P_2(x^{(2)})=P(S_l)\prod_{j=1}^{2}P_j(x^{(j)}|S_l)/\prod_{j=1}^{2}P_j(x^{(j)}) \tag{7}$$

Using the RHS of equation (7) as the prior for calculating the posterior in position 3 gives

$$\frac{P(S_l)\prod_{j=1}^{2}P_j(x^{(j)}|S_l)}{\prod_{j=1}^{2}P_j(x^{(j)})}P_3(x^{(3)}|S_l)/P_3(x^{(3)})=P(S_l)\prod_{j=1}^{3}P_j(x^{(j)}|S_l)/\prod_{j=1}^{3}P_j(x^{(j)}) \tag{8}$$

Continuing on in this fashion through the pth position yields

$$P(S_l)\prod_{j=1}^{p}P_j(x^{(j)}|S_l)/\prod_{j=1}^{p}P_j(x^{(j)})=P(S_l|x^{(1)},\ldots,x^{(p)}) \tag{9}$$

which is the desired result.

Proof of Corollary 1

Proof. First suppose that the new barcode belongs to species l and that $P(S_l|x^{(1)}, \ldots, x^{(p)})$ is strictly decreasing. This means

$$P(S_l|x^{(1)},\ldots,x^{(p)})<P(S_l|x^{(1)},\ldots,x^{(p-1)}) \tag{10}$$

which can be rewritten as

$$P(S_l|x^{(1)},\ldots,x^{(p)})/P(S_l|x^{(1)},\ldots,x^{(p-1)})<1. \tag{11}$$

Because $x^{(1)}, \ldots, x^{(p)}$ are both marginally and conditionally independent given species l, we can expand both numerator and denominator on the LHS as in equation (5). Thus we obtain

$$[P(S_l)\prod_{j=1}^{p}P(x^{(j)}|S_l)/\prod_{j=1}^{p}P(x^{(j)})]/[P(S_l)\prod_{j=1}^{p-1}P(x^{(j)}|S_l)/\prod_{j=1}^{p-1}P(x^{(j)})]<1. \tag{12}$$

On the LHS, the denominator cancels all but the pth terms in the numerator leaving

$$P(x^{(p)}|S_l)/P(x^{(p)})<1 \tag{13}$$

which can be rewritten as

$$P(x^{(p)}|S_l)<P(x^{(p)}). \tag{14}$$

Finally, the RHS can be expanded using the Law of Total Probability to give

$$P(x^{(p)}|S_l)<\sum_{l=1}^{s}P(S_l)P(x^{(p)}|S_l). \tag{15}$$

But $P(x^{(p)}|S_l)$ contains all the mass at position p for species l because the barcodes within a species are identical by assumption. This means the LHS of equation (15) is 1 making the RHS impossible because it violates an axiom of probability. By contradiction we conclude that $P(S_l|x^{(1)}, \ldots, x^{(p)})$ is monotone increasing.

Next, suppose that the new barcode does not belong to species l. Because the species in R have unique barcodes by assumption, and the barcodes within a species are identical, also by assumption, there exists a position p, for which $P(x^{(p)}|S_l)=0$ which implies $P(S_l|x^{(1)}, \ldots, x^{(p)})<P(S_l|x^{(1)}, \ldots, x^{(p-1)})$. Therefore the posterior probability of species l, over all of the positions, is not monotone increasing.

References

Altschul, S., W. Gish, W. Miller, E. Myers and D. Lipman (1990): “Basic local alignment search tool,” J. Mol. Biol., 215, 403–410.
Avise, J. (2000): Phylogeography. The history and formation of species, Cambridge, MA: Harvard University Press.

Clare, E., B. Lim, M. Engstrom, J. Eger and P. Hebert (2006): "DNA barcoding of Neotropical bats: species identification and discovery within Guyana," http://www.barcodeoflife.org/barcode/batsbirds/literature/MEN1657 final.pdf.

Denver, D., K. Morris, M. Lynch, L. Vassilieva and W. Thomas (2000): "High direct estimate of the mutation rate in the mitochondrial genome of caenorhabditis elegans," Science, 289, 2342–2344.

DeSalle, R. (2006): "Species discovery versus species identification in DNA barcoding efforts: response to rubinoff," Conserv. Biol., 20, 1545–1547.

DIMACS (2007): "Center for discrete mathematics and theoretical computer science," http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007.

Dove, C. (2000): "A descriptive and phylogenetic analysis of plumulaceous feather chatacters in Charadriiformes," Ornithological Monographs, 51, 1–163.

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," Infection, Genetics and Evolution, 8, 727–736.

Gascuel, O. and M. Steel (2006): "Neighbor-joining revealed," Mole. Biol. Evol., 23, 1997–2000.

Hajibabaei, M., J. DeWaard, N. Ivanova, S. Ratnasingham, R. Dooh, S. Kirk, P. Mackie and P. Hebert (2005a): "Critical factors for assembling a high volume of DNA barcodes," Philos. Transact. R. Soc. (B), 360, 1959–1967.

Hajibabaei, M., D. Janzen, J. Burns, W. Hallwachs and P. Hebert (2005b): "DNA barcodes distinguish species of tropical Lepidoptera," http://www.pnas.org/content/103/4/968.full.

Hammond, P. (1992): Global biodiversity: status of the Earth's living resources. London: Chapman & Hall.

Hebert, P., A. Cywinska, S. Ball and J. deWaard (2003a): "Biological identifications through DNA barcodes," Proc. R. Soc. (B), 270, 313–322.

Hebert, P., S. Ratnasingham and J. deWaard (2003b): "Barcoding animal life: Cytochomec oxidase subunit 1 divedrgences among closely related species," Proc. Biol. Sciences, 270, S96–S99.

Johns, G. and J. Avice (1998): "A comparative summary of genetic distances in the vertbrates from the mitochondrial cytochrome b gene," Mole. Biol. Evol., 15, 1481–1490.

Karlin, S. and S. Altschul (1990): "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," Proc. Natl. Acad. Sci., 2264–2268.

Kelly, R., I. Sarkar, D. Eernisse and R. DeSalle (2006): "DNA barcoding using chitons (genus Mopalia)," Mole. Ecol. Notes, 7, 177–183.

Kerr, K., M. Stoeckle, C. Dove, L. Weigt, C. Frances and P. Hebert (2007): "Comprehensive DNA barcode coverage of North American birds," http://www.barcodeoflife.org/barcode/batsbirds/literature/MEN1670 final.pdf.

Kimura, M. (1980): "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences," J. Mol. Evol., 16, 111–120.

Koski, L. and G. Goulding (2001): "The closest BLAST hit is often not the nearest neighbor," J. Mol. Evol., 52, 540–542.

Marra, P., C. Dove, R. Dolbeer, N. Dahlan, M. Heacker, J. Whatton, N. Diggs, C. France and G. Henkes (2009): "Migratory canada geese cause crash of US Airways flight 1549," Front. Ecol. Environ., 7, 297–301.

McCallum, A. and K. Nigam (1998): "A comparison of event models for naïve Bayes text classification," Technical Report WS-98-05, AAAI-98 Workshop on Learning for Text Categorization, URL http://www.cs.cmu.edu/ mccallum.

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," Mole. Ecol. Notes, 7, 355–364.

Saitou, N. and M. Nei (1987): "The neighbor-joining method: A new method for reconstruction phylogenetic trees," Mole. Biol. Evol., 4, 406–425.

Stoeckle, M. (2003): "Taxonomy, DNA, and the barcode of life," Bioscience, 53, 2–3.

Studier, J. and K. Keppler (1988): "A note on the neighbor-joining algorithm of Saitou and Nei," Mole. Biol. Evol., 5, 729–731.

Ward, R., T. Zemlak, B. Innes, P. Last and P. Hebert (2005): "DNA barcoding Australia's fish species," http://www.fishbol.org/PDF/ward etal 2005 philtrans.pdf.

Zhang, H. (2004): "The optimality of naive Bayes," in Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference, Miami Beach, FL: The AAAI Press, 562–567.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2013-0025) offers supplementary material, available to authorized users.