

## ACKNOWLEDGMENTS

I thank Phil Cantino, Mike Lee, Jason Anderson, and Kurt Pickett for comments on an earlier version of this paper and Bob O'Hara (ca. 1994) for increasing my own awareness of the distinction between taxonomy and nomenclature.

## REFERENCES

- Cantino, P. D., and K. de Queiroz. 2004. PhyloCode: A phylogenetic code of biological nomenclature. [http://www.ohiou.edu/phylocode/]
- de Queiroz, K. 1992. Phylogenetic definitions and taxonomic philosophy. *Biol. Philos.* 7:295–313.
- de Queiroz, K. 1997. The Linnaean hierarchy and the evolutionization of taxonomy, with emphasis on the problem of nomenclature. *Aliso* 15:125–144.
- de Queiroz, K. 2005. Linnaean, rank-based, and phylogenetic nomenclature: Restoring primacy to the link between names and taxa. *Symb. Bot. Ups.* 33(3):127–140.
- de Queiroz, K., and P. D. Cantino. 2001. Phylogenetic nomenclature and the PhyloCode. *Bull. Zool. Nomencl.* 58:254–271.
- de Queiroz, K., and J. Gauthier. 1990. Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. *Syst. Zool.* 39:307–322.
- de Queiroz, K., and J. Gauthier. 1992. Phylogenetic taxonomy. *Annu. Rev. Ecol. Syst.* 23:449–480.
- de Queiroz, K., and J. Gauthier. 1994. Toward a phylogenetic system of biological nomenclature. *Trends Ecol. Evol.* 9:27–31.
- International Botanical Congress. 2000. International Code of Botanical Nomenclature. Edition adopted by the Sixteenth International Botanical Congress, St. Louis, Missouri, July–August 1999. Koeltz Scientific Books, Königstein.
- International Commission on Zoological Nomenclature. 1999. International Code of Zoological Nomenclature, 4th edition. International Trust for Zoological Nomenclature, London.
- International Union of Microbiological Societies. 1992. International Code of Nomenclature of Bacteria and Statutes of the International Committee on Systematic Bacteriology and Statutes of the Bacteriology and Applied Microbiology Section of The International Union of Microbiological Societies. American Society for Microbiology, Washington.
- Laurin, M., K. de Queiroz, P. Cantino, N. Cellinese, and R. Olmstead. 2005. The PhyloCode, types, ranks, and monophyly: A response to Pickett. *Cladistics* 21:605–607.
- Pickett, K. M. 2005. The new and improved PhyloCode, now with types, ranks, and even polyphyly: A conference report from the First International Phylogenetic Nomenclature Meeting. *Cladistics* 21:79–82.

First submitted 15 April 2005; reviews returned 8 August 2005;

final acceptance 24 August 2005

Associate Editor: Rod Page

*Syst. Biol.* 55(1):162–169, 2006  
Copyright © Society of Systematic Biologists  
ISSN: 1063-5157 print / 1076-836X online  
DOI: 10.1080/10635150500431239

## Statistical Approaches for DNA Barcoding

RASMUS NIELSEN<sup>1</sup> AND MIKHAIL MATZ<sup>2</sup>

<sup>1</sup>Department of Biological Statistics and Computational Biology, Center for Bioinformatics, University of Copenhagen Universitetsparken 15, 2100 Copenhagen, Denmark; E-mail: rn28@cornell.edu

<sup>2</sup>Whitney Laboratory and Department of Molecular Genetics and Microbiology, University of Florida, 9505 Ocean Shore Blvd, Saint Augustine, FL 32080, USA

The use of DNA as a tool for species identification has become known as “DNA barcoding” (Floyd et al., 2002; Hebert et al., 2003; Remigio and Hebert, 2003). The basic idea is straightforward: a small amount of DNA is extracted from the specimen, amplified and sequenced. The gene region sequenced is chosen so that it is nearly identical among individuals of the same species, but different between species, and therefore its sequence, can serve as an identification tag for the species (“DNA barcode”). By matching the sequence obtained from an unidentified specimen (“query” sequence) to the database of sequences from known species, one can thus determine the species affiliation of the specimen. Importantly, the specimen may represent any developmental stage or be just a small fragment of the whole organism, displaying no morphological traits required for standard identification. Although this technique will by no means eliminate the need for the traditional descriptive taxonomy (Dunn, 2003; Lipscomb et al., 2003; Seberg et al., 2003), it is nevertheless envisioned as a key element of future taxonomy research (Stoockle, 2003; Tautz et al., 2003). The

idea of DNA barcoding, although perhaps not surprisingly being a matter of heated debate among dedicated taxonomists (see *Trends in Ecology and Evolution*, volume 18, no. 2, 2003; Will and Rubinoff, 2004), gained rapid acceptance among biologists from other fields. According to the news report in the April 2004 issue of *Nature*, the Barcode of Life Initiative—an international consortium of museums with the secretariat at the National Museum of Natural History in Washington DC—is being established with the goal of creating a database of DNA barcodes from known animal species based on mitochondrial gene cytochrome *c* oxidase subunit I. The DNA barcoding protocol has been already adopted by the Census of Marine Life, a growing global network of researchers in more than 50 countries engaged in a 10-year initiative to assess and explain the diversity, distribution, and abundance of life in the ocean (O’Dor, 2004).

The weakest spot of DNA barcoding is the obvious fact that no gene can serve as an ideal barcode, i.e., be always invariant within species but different among species. It has been pointed out by several authors that

DNA-based identification, if it is to become a rigorous analysis, should be concerned about distinguishing intraspecific from interspecific variation rather than simply recording perfect and imperfect sequence matches (Dunn, 2003; Lipscomb et al., 2003; Stoeckle, 2003). To get around this problem, at present it is assumed that the interspecific sequence variation should exceed a certain threshold, say, 2% or 3% dissimilarity, set on the basis of empirical observations of sequence differences among congeneric species (Hebert et al., 2003a, 2003b). Such an approach seems to be too simplistic to avoid inconclusive or erroneous results. Clearly, there is a need for statistical methods for assessing if a sampled query sequence is sufficiently similar to a particular data base sequence to justify a species assignment of the query.

There are several possible statistical approaches to this problem. In the classical hypothesis-based (frequentist) approach, the null hypothesis could be that the query sequence is a member of a particular species. Such a test would control the rate at which the query is assigned to the true species, i.e., it controls the rate of false negatives. In the Bayesian approach the posterior probability of a species assignment is calculated by assuming a prior distribution of species assignments. We will discuss the problems and merits of these approaches and provide some guidelines towards the use of statistics in DNA barcoding experiments.

#### WHEN SIMILARITY IS MISLEADING

The most obvious source of error is that the database sequence most similar to the query may not be from the species to which the query belongs (the true species). Beside human error (incorrect identification of the specimen used to derive the database sequence), there are at least three reasons why this may happen. First, the true species may not be represented in the database. Second, because of lineage sorting, the random coalescences of lineages in a common ancestral species, the query may not be genetically most closely related to the database sequence from the true species. Third, even if the true species sequence and the query sequence share a most recent common ancestor before either of them share an ancestor with any sequences from other species, the random process at which mutations arise on lineages may cause the sequence representing another species to be most similar to the query. The probabilities that the latter two events occur can be assessed using population genetic theory. For simplicity, we will assume that there are two possible database species, a 'true' and a 'wrong' species. Assuming that both species are panmictic and of constant size, the probability that the query sequence does not share a most recent common ancestor with the true species, before either of them share a common ancestor with the wrong species, is simply  $(2/3)e^{-T/N}$ , where  $T$  is the divergence time between the two species in number of generations and  $N$  is the effective chromosomal population size of the true species (see, e.g., Hudson, 1990). When taking into account the mutational process, and assuming an infinite sites model (e.g., Watterson,

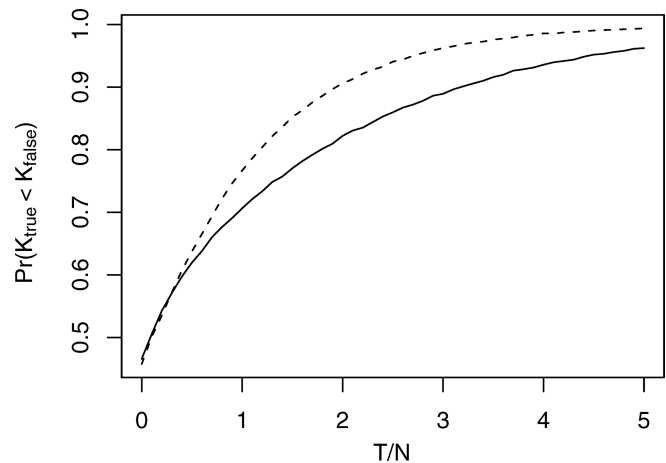


FIGURE 1. The probability that the number of nucleotide differences between the query sequence and another sequence from the same species is smaller than the number of nucleotide differences between the query sequence and a sequence from a different species, as a function of the (scaled) divergence time ( $T$ ) between the two species for  $\theta = 1.0$  (solid) and  $\theta = 5.0$  (dotted line). The probability has been calculated conditional on at least one nucleotide difference between the two database sequences.

1975), calculations can be done numerically, or by simulation to determine assignment probabilities. For example, we might be interested in evaluating the probability that the number of nucleotide differences between the query sequence and the true species sequence is smaller than the number of nucleotide differences between the query sequence and the wrong species sequence [ $\Pr(K_{\text{true}} < K_{\text{false}})$ ]. This probability is shown in Figure 1 assuming that the population size did not change after the speciation of the two species. The probability is calculated for two different values of  $\theta$  ( $= 2N\mu$ , where  $\mu$  is the mutation rate for the entire sequence per generation). To have more than a 95% chance of assigning the sequence to the correct species, the divergence time (scaled by the population size) must be larger than 2.7 when  $\theta = 5.0$  and 4.6 when  $\theta = 1.0$ .

The corresponding probability that the number of nucleotide differences is smallest between the query sequence and the false sequence,  $\Pr(K_{\text{true}} > K_{\text{false}})$ , is shown in Figure 2. Notice that for relatively small mutation rates ( $\theta = 1$ ), the probability of assigning the sequence to the wrong species is larger than 5% until  $T/N$  is larger than 3. These results demonstrate that it is not possible to avoid the problem of within species variability by choosing genes that show very little variability within species, since such genes will tend to provide the lowest discriminatory power. For a given divergence time, it is always better to have as high as possible a mutation rate in the genetic region under investigation. However, for any particular species there may of course be genes that for random reasons are more diagnostic than others.

In the literature, it has been suggested to use a fixed cut-off in terms of pairwise difference for the purpose of species assignment (Hebert et al., 2003a, 2003b). How

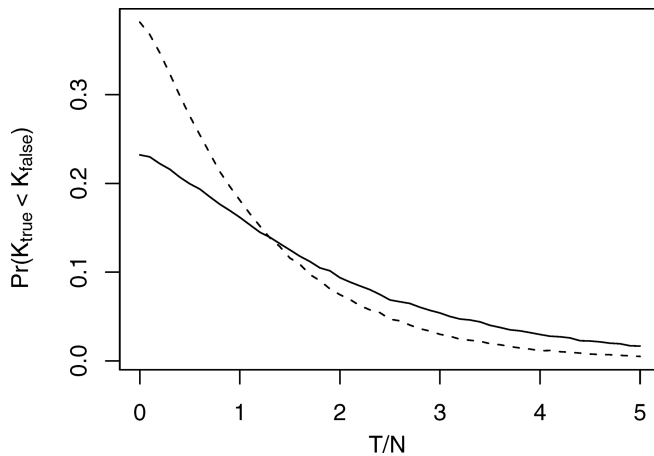


FIGURE 2. The probability that the number of nucleotide differences between the query sequence and another sequence from the same species is larger than the number of nucleotide differences between the query sequence and a sequence from a different species, as a function of the (scaled) divergence time ( $T$ ) between the two species for  $\theta = 1.0$  (solid) and  $\theta = 5.0$  (dashed line). The probability has been calculated conditional on at least one nucleotide difference between the two database sequences.

well such a strategy performs depends on the value of  $\theta$ . Figure 3 shows the probability of observing more than 3, 6, and 9 nucleotide differences between two sequences of the same species as a function of  $\theta$ , calculated using standard methods from coalescence theory. Clearly, the appropriate cut-off depends strongly on the effective population size of the species, indicating that the choices regarding species assignment must take effective population sizes of the relevant species into account.

#### DNA BARCODING USING HYPOTHESIS TESTING

The simplest frequentist approach to consider would be where the null hypothesis is defined as membership in a predefined species. The test procedure should then control the proportion of time the true null hypothesis of species membership is falsely rejected. Here we will explore the properties of such a test using the num-

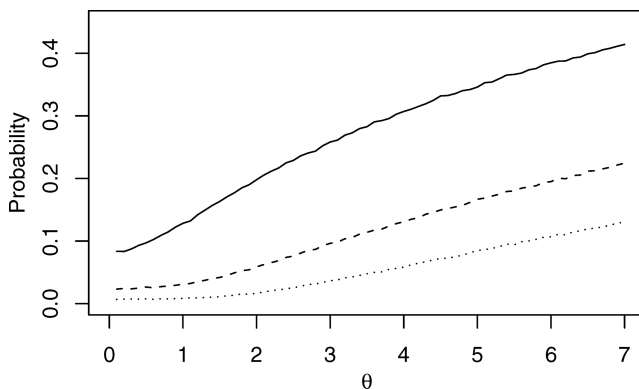


FIGURE 3. The probability of observing more than 3 (solid), 6 (dashed), and 9 (dotted) nucleotide differences between two sequences of the same species as a function of  $\theta$ .

ber of mutational events between the query sequence and the database sequence from the focal species, as a test statistic. Other test statistics could be defined, for example, likelihood ratio tests. However, it should be noted that such tests would have to make more assumptions than the current tests. Also, we will assume an infinite sites model of mutation (Watterson, 1975), although the results easily could be generalized to other models.

First, consider the unrealistic case where  $\theta$  is known with absolute certainty. Then the probability of observing  $K$  nucleotide differences between two sequences of the same species is geometrically distributed with parameter  $1/(1 + \theta)$ , (see, e.g., Hudson, 1990), so

$$\Pr(K \leq k | \theta) = 1 - \left( \frac{\theta}{\theta + 1} \right)^{k+1}, \quad (1)$$

$\theta$  can never be known exactly, so we must represent our knowledge regarding  $\theta$  in a form of a distribution rather than a single value. This distribution should have a fairly flexible functional form and sample space on  $(0, \infty)$ . We will use the gamma distribution, which meets these criteria (although other distributions, such as a truncated normal distribution, could also be considered). Under the assumption of a gamma distribution

$$\Pr(K \leq k) = \int_0^{\infty} \Pr(K \leq k | \theta) f(\theta) d\theta, \quad (2)$$

where

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \quad \theta > 0. \quad (3)$$

and the parameters  $\alpha$  and  $\beta$  must be positive. Although, Equation 2 does not have an analytical solution, it can be evaluated very fast numerically. The cut-off should then be chosen as the minimum value of  $k$  for which  $\Pr(K \leq k) \geq 1 - \alpha$ , where  $\alpha$  is the chosen significance level. The power of this test for  $\alpha = 0.05$  to reject a false null hypothesis is shown in Figure 4, as a function of the divergence time.

This test provides a more rigorous way of determining cut-offs for hypotheses regarding species memberships based on pairwise differences. However, we note that there are many problems associated with this approach, including the fact it does not come with a suitable method for correcting for multiple testing when applied to database searches and that it is based on restrictive assumptions regarding the underlying population genetics. In particular, it relies on assumptions regarding  $\theta$ , and its distribution. If only a single sequence is available from each species, information regarding  $\theta$  is not available. Also, focusing on pairwise differences in the presence of multiple sequences clearly leads to a loss of information.

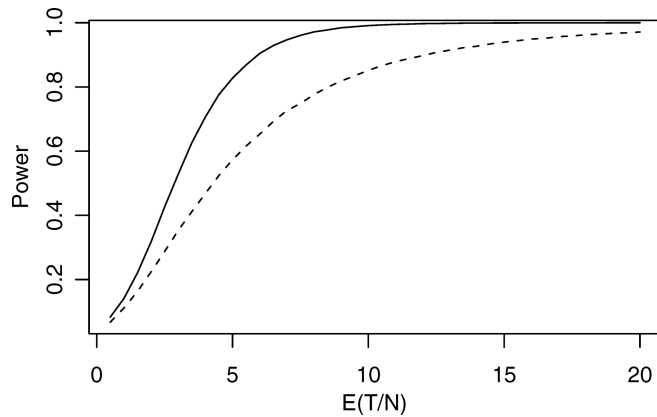


FIGURE 4. The power to reject the membership in the wrong species for different values  $\theta$  and different values of the divergence time  $T/N$  when a species is only represented by a single sequence, for  $\theta = 1$ , (dotted line) and  $\theta = 3$  (solid line). It is assumed that the variance in  $\theta$  equals  $0.25\theta$ .

The previous discussion illustrates that it is not possible to make DNA barcoding inferences without assumptions about  $\theta$ . But what happens if we try to use as little prior information regarding  $\theta$  as possible in DNA barcoding inferences? Mathematically this corresponds to assuming that  $\theta$  is uniformly distributed between 0 and  $m$ , and then considering the limit of  $m \rightarrow \infty$ . In that case  $\Pr(K \leq k) \rightarrow 0$  for any fixed value of  $k$ , suggesting that we should always accept the species assignment. Indeed, any pairwise divergence, however large, may be due to within-species variation caused by a correspondingly large value of  $\theta$ . In other words, DNA barcoding based on single sequences is impossible without implicit or explicit assumptions regarding  $\theta$ . In particular, very large values of  $\theta$  must be excluded a priori.

In practical applications, there are several population genetic tools that can be used to estimate  $\theta$  (Nielsen and Wakeley, 2001; Rannala and Yang, 2003; Wilson and Balding, 1998). For a situation in which only a single sequence has been obtained to represent a species in the database, the easiest solution would be to assume that its  $\theta$  is the same as  $\theta$  in the related species of similar ecology. More sophisticated methods for such an interpolation may be envisioned, which would allow the population size to vary according to some stochastic process along the lineage of a phylogeny. Although such procedures would obviously be of great value for DNA barcoding, they are not the topic of this paper and will not be discussed further. Also, in the case where multiple sequences are known from each species, the problem regarding the true value of  $\theta$  is reduced because the sequences themselves contain information about  $\theta$ .

#### BAYESIAN APPROACH

In some cases a frequentist framework may not be the most suitable for assignment of individuals to species. In particular, when multiple hypotheses are considered at

the same time (i.e., potential membership in many possible candidate species), generating a need for procedures correcting for multiple testing, frequentist procedures based on hypothesis testing may have little power. If we are willing to make assumptions regarding the evolutionary relationships among species, a more natural framework may in some cases be to calculate the posterior probability that the query sequence belongs to any particular database species. Such a Bayesian procedure will be outlined here.

Let  $\mathbf{D} = \{D_1, D_2, \dots, D_k\}$  be the set of database DNA sequences and let  $X$  be the query sequence. We consider  $D_i$  to represent a larger collection of sequences all belonging to the same species, i.e.,  $D_i \in \mathbf{S}_i$ , where  $\mathbf{S}_i$  is the set of mostly unobserved sequences belonging to species  $i$ . We are then interested in assessing the probability of  $X \in \mathbf{S}_i$  given the model of evolution and all the database sequences. Using Bayes' formula and assuming that  $X$  could belong to any species with equal probability a priori, we have

$$\begin{aligned} \Pr(X \in \mathbf{S}_i | \mathbf{D}, X) &= \frac{\Pr(X, \mathbf{D} | X \in \mathbf{S}_i) \Pr(X \in \mathbf{S}_i)}{\sum_{j=1}^k \Pr(X, \mathbf{D} | X \in \mathbf{S}_j) \Pr(X \in \mathbf{S}_j)} \\ &= \frac{\Pr(X, \mathbf{D} | X \in \mathbf{S}_i)}{\sum_{j=1}^k \Pr(X, \mathbf{D} | X \in \mathbf{S}_j)} \end{aligned} \quad (4)$$

In this notation, the presence of nuisance parameters, such as effective population sizes and divergence times has been suppressed. It is important to note that  $\Pr(X, \mathbf{D} | X \in \mathbf{S}_i)$  is not in general proportional to  $\Pr(X, D_i | X \in \mathbf{S}_i)$ —in other words, probabilities of query assignments to different species are not independent—because of evolutionary correlations due to the shared phylogeny. Methods based on approximating  $\Pr(X, \mathbf{D} | X \in \mathbf{S}_i)$  by  $\Pr(X, D_i | X \in \mathbf{S}_i)$  will, therefore, not perform well (results not shown). Establishing a valid Bayesian assignment method for DNA barcoding is, therefore, more difficult than in other apparently similar cases, such as paternity assignment based on unrelated individuals.

Equation 4 can be evaluated using Markov chain Monte Carlo (MCMC) procedures. Importantly for DNA barcoding, established MCMC methods allow explicit integration over the set of nuisance parameters and incorporation of information from multiple sequences from each species. Here, we will modify the procedure of Nielsen and Wakeley (2001) for the purpose of species assignment. In brief, for two species (or populations) this method establishes a Markov chain with state-space given by the set of all possible gene trees, effective population sizes, divergence times, and migration rates between the two species, and stationary distribution proportional to the joint posterior of these parameters. Inferences are then done by simulating this chain and sampling parameter values from the chain at stationarity (when it has reached equilibrium). The distribution of these parameter values then provides an estimate of the posterior density (for further details see Nielsen and

Wakeley, 2001). This procedure can be modified to evaluate Equation (4) by allowing the population identity of an individual to be a random variable, with prior probability of belonging to each species (population) of 50%. To model species divergence, we will assume that the migration rate between the two populations is zero. Uncertainty regarding the population-affiliation of a particular sequence (the query sequence) can then be accommodated by incorporating updates of the affiliation of the sequence into the MCMC scheme. Without migration, such updates will only be accepted when the most recent coalescent between the query sequence and any other lineage in the genealogy occur before the two populations diverge. The update implemented here then simply consists of switching the affiliation of a species without changing the coalescence times or any other aspect of the gene tree. The acceptance probability of such an update is then  $\min\{1, p(G' | \Theta) / p(G | \Theta)\}$ , where  $G'$  is the new proposed gene tree and  $\Theta$  is a vector of population level parameters. At stationarity, the expected proportion of time in the chain the query sequence belongs to species  $i$  is given by Equation 5. Similar schemes could also be used to develop barcoding-aimed procedures on the basis of other methods (Rannala and Yang, 2003; Wilson and Balding, 1998).

The method is currently implemented for the HKY model of DNA sequence evolution and the infinite sites model. At the taxonomic level where barcoding is am-

biguous, the exact model of substitution probably makes little difference.

#### DATA ANALYSIS EXAMPLES

We examined two real data sets that may present a problem for DNA barcoding, representing two marginal cases of extremely low and extremely high sequence variability, both at the intra- and interspecific levels. The first data set contains sequences from the skipper butterfly *Astraptus fulgerator*, which recently has been proposed to be a complex of perhaps as many as 12 separate species (Hebert et al., 2004). Genetic differentiation between these species was originally identified by the phylogeny of cytochrome *c* oxidase subunit I (COI) sequences, and was corroborated by the presence of a morphological difference in caterpillars and the species of plants preferred by them as food. Still, both the degree of divergence within and between these species is very small (Fig. 5A). In sharp contrast to the butterflies, our second example—four species of the Australian rainforest frogs of the genus *Litoria*—displayed intraspecific COI sequence often exceeding 10% pairwise difference (Fig. 5B) (Schneider et al., 1998). High levels of COI variability appear to be common in amphibians (Vences et al., 2005).

To evaluate the power and accuracy of the hypothesis-based test (*K*-test) in application to these two cases,

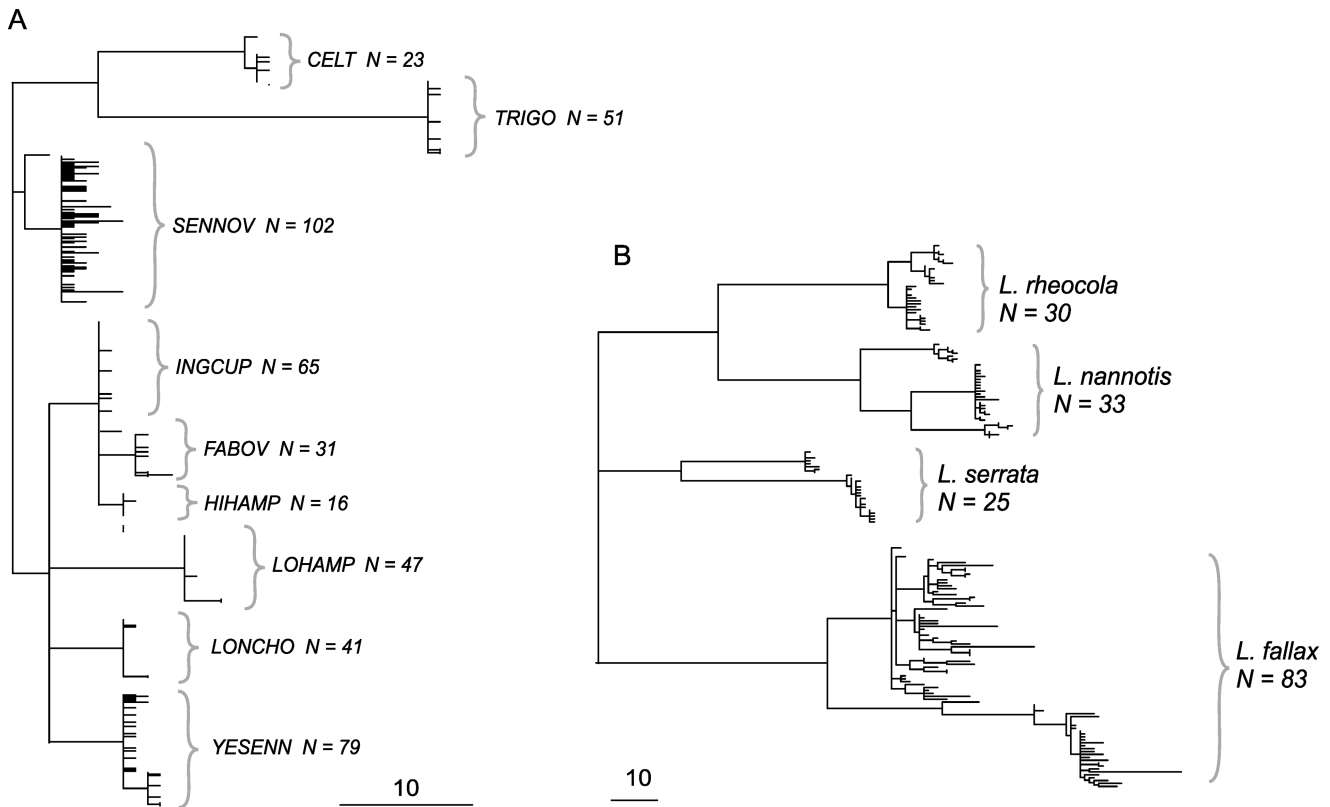


FIGURE 5. Consensus maximum parsimony trees for COI sequences from the two models analyzed here. (A) Skipper butterfly *Astraptus fulgerator* species complex. (B) Four species of the tree frogs of the genus *Litoria*. Scale bars: 10 nucleotide changes. The number of individual sequences per species is indicated near species names.

we performed simulations where in each replicate, “database” sequences (one per species) and one “query” sequence were randomly drawn from the available COI data. These simulated data sets were analyzed using the values of  $\theta$  and their variances estimated from the complete collection of sequences for each species. For each species, we performed 100 replicates. When the query sequence matched the database with a  $P$ -value exceeding 0.05 for one and only one species, the identification was designated “correct” if the query sequence was indeed derived from that species, and would be “wrong” if not. If the test sequence matched more than one database species, it was considered a “tie.” Occasionally, the query sequence did not match any of the database species with sufficient  $P$ -value, which was designated “no ID.” We also performed the same analysis using identical values of the mean and variance of  $\theta$  for all species, which were equal to their averages over all the analyzed species. This analysis was intended to approximate a situation where no population genetic data are available for one or more species in the database, and their values of  $\theta$  are interpolated using data from related species of similar ecology.

The results of the  $K$ -test simulations are summarized in Table 1. Notably, we never encountered a case of wrong identification. Using the experimental values of  $\theta$ , of the nine tested groups within the *Astrartes fulgerator* complex, four could be identified with very high power (0.99–1) at the 0.05 significance level, while another three were identifiable with the power 0.83–0.91. The power for the two remaining groups—INGCUP and HIHAMP—was 0.33 and 0.25, respectively, due to frequent ties between themselves and also with the third very closely related FABOV group (Fig. 5). LOHAMP or LONCHO queries, although matching the database sequence of their own groups with high  $P$ -value, in about 10 percent of cases matched the YESENN database sequence with  $P$ -value slightly exceeding 0.05 due to higher genetic diversity within YESENN, thereby generating ties. The great number of ties observed for *Litoria* species was

due to a similar situation: in the majority of cases the query matched the species with highest estimate of  $\theta$  (*L. fallax* and/or *L. serrata*) with  $P$ -value that exceeded 0.05 slightly, in addition to providing a good match to the correct species. No “no ID” cases were observed during *Litoria* simulations. Interestingly, the use of the average  $\theta$  and variance for all competing species resulted in 100% correct identification in *Litoria* and not much change of power for *Astrartes fulgerator*, except for FABOV group (Table 1), suggesting that in these cases, interpolation of  $\theta$  from related species may be a reasonable option.

We then investigated whether a Bayesian method could help in resolving the ties such as the ones suggested by the  $K$ -test. Because the current version of the method is able to discriminate between only two species, we focused on the two pairs from our examples that were represented by more than 20 sequences each and produced most ties: INGCUP and FABOV groups of the butterflies and *L. fallax* and *L. serrata* species of the frogs. We generated 100 datasets that each included one, three, or five randomly chosen database sequences from the two competing species plus one test sequence from FABOV or *L. serrata*. This data was then analyzed assuming a truncated uniform prior for  $\theta$  ( $U[0, 10]$ ) and the scaled divergence time  $t$ .

The accuracy of the Bayesian sorting at the 0.95 confidence level exceeded 0.9 both for butterflies and frogs when three or five database sequences were used, being considerably less for one database sequence (Fig. 6).

We also studied the behavior of the Bayesian sorting when the query sequence was not derived from one of the two database species (*L. fallax* and *L. nannotis*), but from the third one (*L. serrata*). In these simulations (100 replicates), the database for each of the two species included one or three sequences. These experiments mimic the case when the query comes from an “unrecorded” species. In such a situation the  $K$ -test would tend to assign the query to the species with the highest value of  $\theta$ —*L. fallax*. In contrast, in Bayesian sorting the posterior

TABLE 1. Results of the  $K$ -test simulations.

	$\theta^a$	SD	Correct	Experimental $\theta$ /SD Tie	No ID	Correct	Average $\theta$ /SD Tie	No ID
Skipper butterfly <i>Astrartes fulgerator</i> complex:								
CELT	0.91	0.73	100	0	0	97	3	0
TRIGO	0.31	0.36	100	0	0	100	0	0
SENNOV	1.45	0.99	99	0	1	92	0	8
INGCUP	0.15	0.24	33	65	2	1	99	0
FABOV	0.7	0.61	91	7	2	22	76	2
HIHAMP	0.13	0.23	25	75	0	1	99	0
LOHAMP	0.36	0.39	83	8	9	99	1	0
LONCHO	0.27	0.33	89	11	0	100	0	0
YESENN	1.59	1.06	100	0	0	99	1	0
Average	0.65	0.55						
Tree frogs <i>Litoria</i> :								
<i>L. fallax</i>	38	19	81	19	0	100	0	0
<i>L. serrata</i>	31	16	0	100	0	100	0	0
<i>L. nannotis</i>	25	12	5	95	0	100	0	0
<i>L. rheocola</i>	11	6	8	92	0	100	0	0
Average	26	13						

<sup>a</sup>  $\theta$  was estimated as nucleotide diversity per locus.

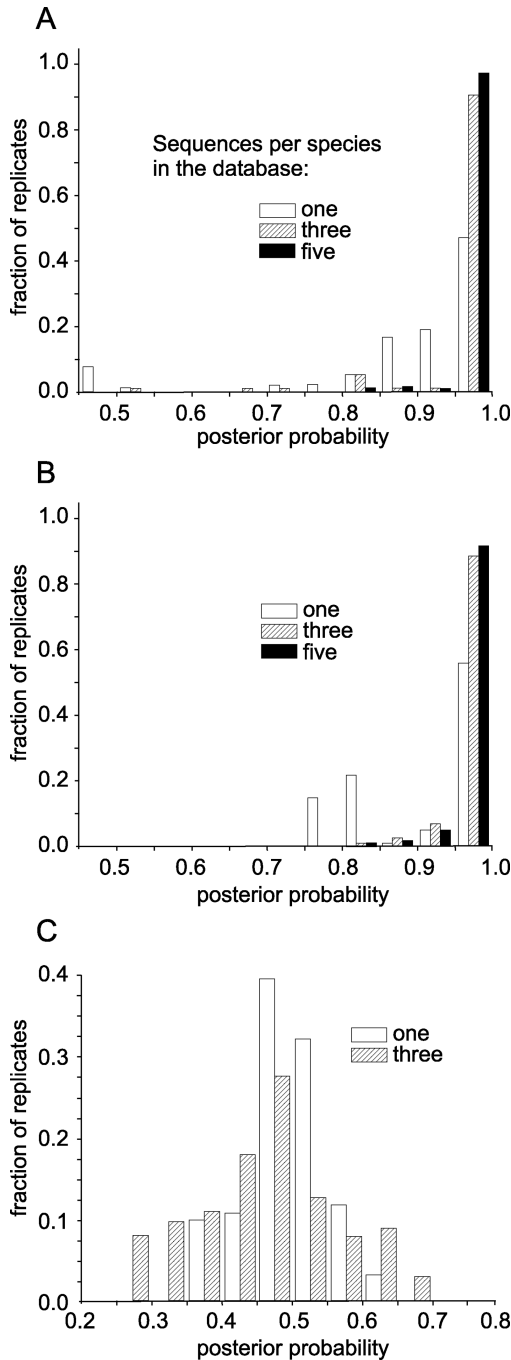


FIGURE 6. Histograms of posterior probabilities obtained in Bayesian assignment simulations, for different number of sequences per species in the database. (A) *A. fuligator*, FABOV versus INGCUP; the query sequence comes from FABOV. (B) *Litoria fallax* versus *Litoria serrata*, the query sequence comes from *L. serrata*. (C) *Litoria fallax* versus *Litoria nannotis*, the query sequence comes from *L. serrata* ("unrecorded species"). X-axis: posterior probabilities (bins of 0.05); Y-axis: fraction of replicates (out of 100).

probabilities were grouped around 0.50 and were never less than 0.25 or more than 0.70, rightfully suggesting that the query sequence does not belong to any of the database species (Fig. 6). The probabilities were more

tightly grouped around 0.5 when one database sequence was used, reflecting the tendency seen on Figure 6A and B: the method is reluctant to assign a query to any species when the database is small.

#### DISCUSSION

Once the unrealistic assumption of perfect sequence identity within species is abandoned, it becomes clear that DNA barcoding cannot be performed without making population genetic assumptions. Establishing measures of statistical uncertainty in DNA barcoding has to rely on strong assumptions regarding the population genetics of the analyzed species. Still, the inference procedures based on the number of pairwise sequence differences should often be relatively robust to the population genetic assumptions. Most violations of assumptions regarding demography will simply change the (coalescence) effective population size, and only have a minor effect on the distribution of the test statistic. However, it is important to note that there can be cases, particularly cases with strong population subdivision within species, where species assignment may fail because the underlying demographics have not been modeled adequately. An extreme case is a sequence sampled from a sub-population that has no ongoing gene-flow with any of the populations from which the database sequences have sampled. The statistical method discussed here will then be likely not to categorize the query sequence as a member of the focal species, even if taxonomists would recognize it as belonging to it. In this case DNA barcoding may fail because the recognized taxonomic units do not correspond to populations that are reproductively isolated.

Here we have primarily focused on the number of nucleotide changes ( $K$ ) as a statistic in the hypothesis-based (frequentist) approach. Our results suggest that a procedure that examines the number of nucleotide changes only between a query sequence and its best match in the database is not optimal. More accurate methods can be established by simultaneously considering the number of changes between other divergent species in the database and the query sequence. Ultimately, DNA barcoding methods should simultaneously consider the phylogenetic information from all the (relevant) species in the database (be based on the species tree), and the population genetic information available from multiple sequences (incorporate information regarding the coalescent process within species). The Bayesian method presented here is a first step in this direction. Still, it is highly desirable to modify it to account for possible unsampled species, to deal with an arbitrary number of species and to deal with uncertainty regarding  $\theta$  for species represented by none, or very few, sequences.

The  $K$ -test and Bayesian assignment method presented here are the first methods developed thus far to approach DNA barcoding from a perspective of statistical population genetics. Overall, the performance of

the  $K$ -test was good for the case of *Astraptus fulgerator* complex, but less so for the *Litoria* species, where the high values of  $\theta$  in *L. fallax* and *L. serrata* produced many "ties." Given a complete database of barcodes, this could be viewed as a conservative behavior, because the test is reluctant to assign a query to one particular species in the view of the great intraspecific variation observed in the *Litoria* genus. However, if some species were missing from the database, such a behavior could result in incorrect assignment of queries derived from these "unrecorded" species.

The Bayesian approach described here has obvious advantages over the  $K$ -test in terms of accuracy and ability to deal adequately with more than one sequence per species in the database. Furthermore, this method directly integrates over possible values of  $\theta$  alleviating the need to provide estimates of  $\theta$  for each species. As mentioned above, we believe that a generalized version of this approach would be the method of choice for DNA barcoding in the future. However, we also note that this method is not without problems, particularly in making phylogenetic assumptions, and species level assumptions that may not always be correct. Furthermore, the availability of a full Bayesian approach may not eliminate the need for inference procedures with controlled frequentist properties. The computational machinery used in devising the Bayesian approach could also be used in the development of frequentist approaches. Such methods may be used to devise better statistics incorporating more of the information in the data, while appropriately taking uncertainty in the nuisance parameters, such as species divergence times, into account.

The interrelationship between classical taxonomy and DNA barcoding has been a matter of extensive debates. In our view, the situation is quite simple. DNA barcoding can undoubtedly be a useful tool to assign unknown specimens to predefined groups (i.e., species as defined by a classical taxonomy), but never a method for identifying these groups in the first place. The fact that it is possible to assign a sequence from a specimen to a predefined group does not mean that the group itself deserves taxonomic recognition. For example, for *Astraptus fulgerator*, one may define the whole collection of sequences as one group, or separate it into a dozen smaller phylogenetic clades and define these units as species within a species complex. In both cases, statistical approaches, such as the ones described here, could be successfully used to assign query sequences to taxonomic units. When encountering a query that cannot be assigned to any of the groups recorded in the database, one should not automatically infer that this query represents a new species, because the possibility will always remain that one or more of the database species were undersampled. A taxonomic decision concerning such a query should be made on the basis of independent lines of evidence.

#### ACKNOWLEDGEMENTS

Funding for this project was provided by grants from NIH (GM066243) and US Department of Defense (SERDP program) to MVM, and HFSP grant RGY0055/2001-M, NSF grant DEB-0089487, and NSF/NIH grant DMS/NIGMS-0201037 to RN.

#### REFERENCES

- Dunn, C. P. 2003. Keeping taxonomy based in morphology. *Trends Ecol. Evol.* 18:270–271.
- Floyd, R., E. Abebe, A. Papert, and M. Blaxter. 2002. Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11:839–850.
- Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003a. Biological identification through DNA barcodes. *Proc. R. Soc. Lond. B.* 270:313–321.
- Hebert, P. D. N., E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptus fulgerator*. *Proc. Natl. Acad. Sci. USA* 101:14812–14817.
- Hebert, P. D. N., S. Ratnasingham, and J. R. de Waard. 2003b. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. S B Biol. Sci.* 270:S96–S99.
- Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pages 1–44 in *Oxford surveys in evolutionary biology* (D. J. F. A. J. Antonovics, ed.). Oxford University Press, Oxford.
- Lipscomb, D., N. Platnick, and Q. Wheeler. 2003. The intellectual content of taxonomy: A comment on DNA taxonomy. *Trends Ecol. Evol.* 18:65–66.
- Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics* 158:885–896.
- O'Dor, R. 2004. A census of marine life. *Bioscience* 54:92–93.
- 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.
- Remigio, E. A., and P. D. N. Hebert. 2003. Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Mol. Phylogenet. Evol.* 29:641–647.
- Schneider, C. J., M. Cunningham, and C. Moritz. 1998. Comparative phylogeography and the history of endemic vertebrates in the Wet Tropics rainforests of Australia. *Mol. Ecol.* 7:487–498.
- Seberg, O., C. J. Humphries, S. Knapp, D. W. Stevenson, G. Petersen, N. Scharff, and N. M. Andersen. 2003. Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends Ecol. Evol.* 18:63–65.
- Stoeckle, M. 2003. Taxonomy, DNA, and the bar code of life. *Bioscience* 53:796–797.
- Tautz, D., P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler. 2002. DNA points the way ahead of taxonomy—In assessing new approaches, it's time for DNA's unique contribution to take a central role. *Nature* 418:479–479.
- Tautz, D., P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler. 2003. A plea for DNA taxonomy. *Trends Ecol. Evol.* 18:70–74.
- Vences, M., M. Thomas, A. van der Meijden, Y. Chiari, and D. R. Vieites. 2005. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Front. Zool.* 2:5.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7:256–276.
- Will, K. W., and D. Rubinoff. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics Soc.* 20:47–55.
- Wilson, I. J., and D. J. Balding. 1998. Genealogical inference from microsatellite data. *Genetics* 150:499–510.

First submitted 17 January 2005; reviews returned 3 May 2005;

final acceptance 30 August 2005

Associate Editor: Paul Lewis