Testing the Reliability of Genetic Methods of Species Identification via Simulation

HOWARD A. ROSS, SUMATHI MURUGAN, AND WAI LOK SIBON LI
Bioinformatics Institute, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand; E-mail: h.ross@auckland.ac.nz (H.A.R.)

Abstract—Although genetic methods of species identification, especially DNA barcoding, are strongly debated, tests of these methods have been restricted to a few empirical cases for pragmatic reasons. Here we use simulation to test the performance of methods based on sequence comparison (BLAST and genetic distance) and tree topology over a wide range of evolutionary scenarios. Sequences were simulated on a range of gene trees spanning almost three orders of magnitude in tree depth and in coalescent depth; that is, deep or shallow trees with deep or shallow coalescences. When the query’s conspecific sequences were included in the reference alignment, the rate of positive identification was related to the degree to which different species were genetically differentiated. The BLAST, distance, and liberal tree-based methods returned higher rates of correct identification than did the strict tree-based requirement that the query was within, but not sister to, a single-species clade. Under this more conservative approach, ambiguous outcomes occurred in inverse proportion to the number of reference sequences per species. When the query’s conspecific sequences were not in the reference alignment, only the strict tree-based approach was relatively immune to making false-positive identifications. Thresholds affected the rates at which false-positive identifications were made when the query’s species was unrepresented in the reference alignment but did not otherwise influence outcomes. A conservative approach using the strict tree-based method should be used initially in large-scale identification systems, with effort made to maximize sequence sampling within species. Once the genetic variation within a taxonomic group is well characterized and the taxonomy resolved, then the choice of method used should be dictated by considerations of computational efficiency. The requirement for extensive genetic sampling may render these techniques inappropriate in some circumstances. [BLAST; DNA barcoding; phylogenetic; species identification.]

Molecular genetic techniques for species identification based on single-gene sequence similarity or phylogenies are rapidly gaining wide use. A group of techniques, broadly termed “DNA barcoding” (Hebert et al., 2003), use either the mitochondrial cytochrome oxidase I (COI) gene for animals or one of a range of genes for plants (Kress et al., 2005; Little and Stevenson, 2007; Rubinoff et al., 2006a). The species identity of a query (unknown) sequence is assigned on the basis of its similarity to a set of reference (identified) sequences.

Despite the establishment of the Consortium for the Barcode of Life (www.barcoding.si.edu) to oversee the development of reference databases and analysis methods, the actual identification protocols for barcoding remained ambiguous (Rubinoff et al., 2006b) until very recently (Ratnasingham and Hebert, 2007). Consequently, the term “barcoding” has been applied widely to include the identification, delimitation, and description of species by the use of genetic distance comparisons, hierarchical clustering, distance thresholds, and other methods based on single gene comparisons (e.g., Little and Stevenson, 2007). The barcoding method has now (Ratnasingham and Hebert, 2007) been explicitly defined as assigning an identity based on the reference with the minimum genetic distance from the query.

The methods for barcoding have been criticized (Meyer and Paulay, 2005; Moritz and Cicero, 2004; Will and Rubinoff, 2004) for (1) failing to recognize that there is no conspecific in the reference set; (2) assuming that intraspecific variation is well sampled in the reference set; (3) assuming that species are reciprocally monophyletic; (4) assuming that hybrid introgression has not occurred; (5) assuming that reference sequences are correctly identified; (6) assuming that the gene phylogeny estimation is robust; and (7) diverting resources from other areas of systematic research (Cameron et al., 2006).

To address the first concern, researchers have implemented thresholds beyond which a query sequence is considered unassigned (Hebert et al., 2004). A problem with thresholds is that no single sequence dissimilarity threshold can apply to all cases because of variation in species’ levels of intraspecific divergence (Will and Rubinoff, 2004). However, relative thresholds remain a viable approach to enhance a method’s ability to detect when the answer must be ambiguous, and they are explored in this study. In addition, topological thresholds can be employed as an alternative. The second criticism applies in particular to assessment of confidence in species assignment and is expected to vary among methods. The third concern, species paraphyly, is a real biological condition (Funk and Omland, 2003) that is also expected to affect different methods variously and will be studied here. The fourth concern, hybrid introgression, poses a problem for all single-gene methods because the gene genealogy does not match that of the species (Monaghan et al., 2006; Nelson et al., 2007; Whitworth et al., 2007). The seriousness of introgression will depend on the frequency or recentness of its occurrence and is beyond the scope of this study. The last three concerns are outside the scope of the current study—inaccurate taxonomy and poor tree estimation are general problems, and diversification of resources is not a scientific problem. The ambiguous use of the term “barcoding” and the dual use of several techniques in both identification and classification (e.g., Blaxter, 2004; Hebert et al., 2003; Tautz et al., 2003) have contributed to the controversy.

Our goal is to assess the reliability of several different methods of species identification (BLAST, genetic distance, and tree-based methods) that previously were combined under the “barcoding” rubric. Unlike morphology-based methods of identification, which can refer to the diagnostic characters of type specimens.
to define the species, molecular methods are heuristics that can only allow us to make probabilistic statements of species identity. We wish to know under what situations each of these heuristic methods fails to make a correct identification. BLAST is commonly used as an informal method of identification (e.g., Holder and Lewis, 2003; 275), by searching the NCBI reference database (www.ncbi.nlm.nih.gov/BLAST/) for sequences that give the best alignments to all or a part of the query sequence. BLAST has been found to be inconsistent and unreliable (Agarwal and States, 1998; Anderson and Brass, 1998; Woodwark et al., 2001) in certain real-life situations. Little and Stevenson (2007) found that BLAST can give accurate identification at the genus, but not the species, level in gymnosperms, perhaps because top hits are often not closest phylogenetic relatives (Koski and Golding, 2001). Simple measures of genetic distance have commonly been used in barcoding studies to infer identity, with thresholds used to distinguish conspecifics from heterospecifics (Hebert et al., 2003; Ratnasingham and Hebert, 2007). Tree-based analyses were used by Hebert et al. (2003) to justify barcoding and have been used in conjunction with well-curated reference data sets to provide species identifications of cetaceans (Baker et al., 1996; Baker and Palumbi, 1994; Dalebout et al., 1998; Ross et al., 2003; Ross and Murugan, 2006). Will and Rubinoff (2004), among others, have asserted that identity cannot be assigned to a query placed in an ambiguous position on a cladogram, sister to a species clade, but we will subject this to testing.

Statistically more sophisticated methods of species identification have recently been developed. Matz and Neilsen (2005) described a likelihood-ratio test (LRT) to test query species identity. Nielsen and Matz (2006) and Abdo and Golding (2007) used coalescent-based approaches in the context of Markov chain Monte Carlo (MCMC) to assign unknowns to identified species. These methods have not been investigated here because software implementations were not publicly accessible when the study was performed. They are computationally demanding and, in the case of the LRT, corrections for multiple testing remain uncertain. Although these methods are not panaceas, and cannot overcome the biological issues of paraphyly and introgression, they offer potential improvements over previously used methods.

Here we report on simulations designed to assess the relative powers of molecular identification techniques in recovering species identity under different combinations of species birth rates and coalescent depth. The degree to which the rate of species diversification differs from the background rate of neutral DNA sequence substitution is expected to affect the success of these identification methods. Relatively short time intervals between speciation events will not allow sufficient time for lineage sorting or the accumulation of lineage-specific substitutions, thereby reducing accuracy. The mean depth on the phylogenetic tree at which two genes have a common ancestor, the coalescence time, is equal to theta ($\theta$), the long-term effective population size scaled by substitution rate. Deep coalescences, indicative of large or highly structured populations, can be expected to result in increased paraphyly and reduced accuracy in species identification.

Our goal is to test several genetic species identification methods across a broad range of evolutionary scenarios. Recent tests of the barcoding methods (Little and Stevenson, 2007; Meier et al., 2006; Meyer and Pauly, 2005) have been based on a specific well-sampled group of organisms and a single or few genes. The ability to discriminate species identity will be influenced by the incidence of paraphyly, as determined by the depth of gene coalescence within a species relative to the level of genetic divergence between species. By using simulation methods and a range of population and phylogenetic parameters, we survey a wider tree space to investigate under what phylogenetic circumstances we can expect these methods to succeed or fail. Overall, relatively simple models of speciation and sequence evolution were used because we aimed to assess the general usefulness of the identification techniques rather than their success with any particular gene (e.g., COI) or genome (e.g., mitochondrion). The results can therefore be considered “best case,” with the additional complexities of rate variation across sites, correlated rate variation across lineages, highly variable effective population sizes, insertion/deletion events, etc., all likely to reduce the efficacy of these methods.

**Materials and Methods**

Reference sequence alignments and query sequences were simulated across a broad range of phylogenetic trees using a pipeline of readily available software. Sequence data sets of size similar to that of a taxonomic genus or family were simulated under different evolutionary scenarios, from which reference and query sequences with known identity were derived. Then BLAST, genetic distance, and tree-based methods were used to assign an identity to each query sequence and the frequency with which each method returned the correct identification was determined. We consider both complete and incomplete taxon sampling by including or excluding the reference sequences from the species to which the unknown truly belongs.

**Species Tree and Gene Tree Simulations**

First a species tree (S) was simulated using Phylogen v1.1 (Rambaut, 2003) with a constant rate of species birth, and no species death, until a fixed number (50) of extant species was reached. Trees were simulated using several species birth rates (Fig. 1a). Then, a single ultrametric gene tree (T) was simulated using MCCcoal v1.1 (Rannala and Yang, 2003; Yang, 2005) on that species tree S. For each species on the tree, gene coalescences were simulated for a number of gene copies drawn randomly from the range 2 to 20. Gene trees were simulated using several values of theta (Fig. 1a). All of the species had the same value of theta ($\theta$) in any given simulation. In a small number of simulations, each species had a value of theta drawn from a Gaussian distribution with mean...
θ and standard deviation 0.2θ, which resulted in theta varying over a threefold range of values. A nonultrametric gene tree was then obtained by introducing rate heterogeneity across lineages. Rate heterogeneity was simulated by multiplying the length of each branch in the ultrametric gene tree (T) by \(1 + ax\), where x was a value drawn at random from a standard exponential distribution (\(P(x > k) = \exp(-k)\)). The constant a, the rate heterogeneity factor, adjusts the deviation intensity from the molecular clock hypothesis (Guindon and Gascuel, 2002; Makarenkov and Legendre, 2004). Most simulations were performed using \(a = 0.7\), but a small number were performed using 0.5 and 0.9. The monophyly of each species on the gene tree was determined.

Sequences were simulated on non-ultrametric gene trees using Seq-gen v1.3.1 (Rambaut and Grassly, 1997), with the HKY model of evolution, transition/transversion ratio = 3, ACGT nucleotide frequencies = (0.3, 0.2, 0.2, 0.3), and sequence length = 500. A more complicated model of sequence evolution, resembling that observed in COI sequence variation, was tested to assess the influence of the simulation parameters on the outcome. The parameters of the GTR+G+I model (GTR rate matrix = (1.5, 50, 3, 2, 75, 1), nucleotide frequencies = (0.27, 0.21, 0.19, 0.33), proportion invariant sites = 0.5, gamma-distributed rate variation with 4 site categories and shape parameter = 1) were estimated using PHYML (Guindon and Gascuel, 2003) for an alignment of COI sequences from the dog family (Canidae) and applied in a small number of simulations. Seq-gen produces an alignment without gaps, so a subsequent alignment process was not needed. For each species in the sequence alignment, the first sequence was put aside to be used only as a query sequence in the method assessments, and the remaining sequences (1 to 19 per species) became the reference alignment. The true species identity of each query sequence was retained.

For each phylogenetic scenario investigated (Fig. 1a), 100 replicate trees, each including 50 species, were simulated. Because the number of sequences (gene copies) per species was a random variate, there was some small variation between replicates in sample number, but on average 501 (range 376 to 639) sequences were simulated per tree, of which 50 were used as queries and the remaining 451 comprised the references.

**Testing Methods of Genetic Species Identification**

The different methods of species identification were assessed by attempting to use them to identify each of the query sequences in combination with the corresponding reference alignment. This assessment was carried out twice, once with reference sequences for the true species present in the reference alignment and again with those reference sequences removed. In any identification attempt there are three possible outcomes: a correct ID, an incorrect ID (both positive IDs), and an ambiguous ID. Here we make a finer distinction among these outcomes because we are interested in the contributing factors. The first distinction is between positive and ambiguous IDs. A positive ID can be correct only when the true species is in the reference alignment. We term these true- or correct-positive IDs. A positive ID can be incorrect for two very different reasons. In the first case, if the true species was in the reference alignment but the method failed to recognize it, then it is an incorrect positive ID. The second way in which an incorrect ID can arise is if the species is not in the reference alignment and the method failed to return an ambiguous result, giving a false-positive ID. There are two ways in which ambiguous IDs can arise. When the true species is not represented among the references, and no identification is appropriate, then an uncertain result is a true-negative ID. However, when the true species is represented in the reference alignment and a correct identification is possible, but the method is too conservative, then the result is called a false-negative ID. Although there are only three possible outcomes, there are five ways in which they can arise. This nonstandard
terminology has been used so that the strengths and weaknesses of the methods can be better understood. Several variants of the three major methods of identification were assessed.

**BLAST.**—BLAST output was analyzed in four different ways to emulate the intuitive or unconscious decision processes thought to be used by researchers when making species identifications.

- **BLAST1:** the ID is that of the species associated with the best BLAST hit, and E-value < cut-off. This corresponds to choosing the top hit in the BLAST results.
- **BLAST2:** the ID is that of the species having the highest score, where each hit in the top n hits, based on its rank, contributes a score of \((n - \text{rank} + 1)\) to the associated species, given E-value < cut-off, \(n = 10\). This method weights the species in the top 10 hits inversely with their ranking so that a species with several lower quality hits could be selected instead of a single high-ranking hit.
- **BLAST3:** the ID is that of the species with the greatest number of hits having a bit score of at least B% of the top hit, given E-value < cut-off, \(B = 90\). This corresponds to the most common species with a high bit score.
- **BLAST4:** the ID is that of the species having the highest score, where each hit in the top n hits contributes its bit score to its associated species, given E-value < cut-off, \(n = 5\). This combines score quality and number of sequences among the top five hits.

A local installation of BLAST was used to search for the query sequence among the reference sequences. In each case a liberal \((10^{-2})\), or more restrictive \((10^{-6})\), E-value cut-off was used. The query’s ID was uncertain when no hits had an E-value below this cut-off.

**Distance method.**—For each reference alignment, all pairwise genetic distances were computed among the reference sequences, and between each query and each of the reference sequences, using the Kimura two-parameter model of evolution (as used by Hebert et al., 2003). Genetic distances were estimated using routines in the JAVA-language PAL phylogenetics software library (Drummond and Strimmer, 2001), which gives a maximum distance of 1, representing substitutional saturation. Identifications were assigned in two ways.

- **Distance (mean):** the ID is that of the species having the smallest mean genetic distance from the query, given that it is less than the distance threshold.
- **Distance (nearest):** the ID is that of the species having the smallest genetic distance from the query, given that it is less than the distance threshold.

Distance thresholds were used to distinguish conspecifics from heterospecifics. Fixed thresholds have been widely used: Ratnasingham and Hebert (2007) report that the Barcode of Life system now uses a 1% threshold, whereas Abdo and Golding (2007) recently used 3%, as was common in earlier barcoding studies. Here a variable threshold was computed based on the intraspecific variation in the data set. A liberal threshold was obtained by using the 99th percentile, and a more restrictive threshold by using the 90th percentile, of intraspecific distances. The distance (nearest) method corresponds to Meier and colleagues’ (2006) “Best close match,” except that they used the 95th percentile of intraspecific distance as a threshold. An uncertain ID arose when the distance to the nearest sequence or species equalled or exceeded the distance threshold.

**Tree methods.**—The ID of the query sequence was determined by its placement on a neighbor-joining tree (Saitou and Nei, 1987), estimated using genetic distances estimated with the same HKY model and parameters as were used initially to simulate the sequence evolution (Drummond and Strimmer, 2001). Two topological assessments were used to assign an identification.

- **Liberal tree-based: when the query sequence (Q) is either sister to ((X, X), Q), or within ((X, Q), X) a monospecific clade, the ID is that of the species in the clade (X); otherwise, it was considered uncertain. This differs from Meier and colleagues’ (2006) “Tree-based identification, sensu Hebert” (Hebert et al., 2003) where monophyly of the entire species is required.
- **Strict tree-based: when the query sequence is within a monospecific clade ((X, Q), X), the ID is that of the species in the clade; otherwise, it was considered uncertain. This approach is equivalent to Meier and colleagues’ (2006) “Tree-based identification, revised criteria.” Long-branch attraction might result in the misplacement of a query on a tree. To protect against this, the tree-based methods can be used with a threshold.
  - **Liberal tree-based (+threshold) and Strict tree-based (+threshold):** as above but requiring that the nearest reference sequence is less than a distance threshold; otherwise, it was considered uncertain.

Two thresholds were used: the liberal 99th percentile of the intraspecific distances and the more restrictive 90th percentile. In the strict tree-based methods, there was no requirement for all members of the species to be in the clade, but simply that all members of the clade must be of the same species. Because of the strictness of this method, the results were screened to remove cases (~5%) where a species was represented by only one sequence in the reference data set. The results presented have been recomputed using this slightly smaller subset.

**Statistical tests.**—The relative performances of the different methods were compared using the proportion of outcomes (i.e., correct ID, incorrect ID, etc.) per replicate as the observation. The mean proportions for each method were compared within a set of tree parameters using the Kruskal-Wallis rank sum test as implemented.
in the JMP software v5.1. Given the large number of comparisons made, a de facto Bonferroni correction for multiple comparisons was made by choosing \( P = 0.0001 \) as the critical value in all tests.

**Comparison of simulated and empirical trees.**—Computable parameters were required so that the simulated gene trees could be compared with those estimated from empirical sequences. Given that the species birth process in the simulation was uniformly random with respect to lineage, the root of the simulated species and gene trees, on average, was at the middle of a relatively symmetrical tree. Increasing the species birth rate resulted in the target number of species being achieved more quickly, with a resultant shorter tree than would be obtained with a slower species birth rate. The depth of the tree was calculated as the average distance from each tip to the root. The coalescence time (time to most recent common ancestor) for each pair of genes within a species naturally increased as theta increased. An estimate of coalescence time was obtained by computing the average intraspecific distance on the tree (Fig. 1b). Empirical and simulated trees were also compared using a statistic that reflected species differentiation: the average distance between the members of one species and a second species that was on average closest to the first. Because this computation took a long time, it was performed for only 10 randomly chosen gene trees from each evolutionary scenario (Fig. 1c).

For comparison with the simulated trees, phylogenetic trees were estimated using PHYML (Guindon and Gascuel, 2003) and the HKY model of evolution from the following empirical data sets:

1. A reference alignment of mtDNA control region sequences from Odontocete toothed whales (63 species) (Ross et al., 2003).
2. Seven data sets of mtDNA COI sequences from the Barcode of Life Data Systems site (http://www.barcodinglife.org/views/projectlist.php) comprising crustaceans (58 species), fungi (13, 43, and 52 species), *Euglena* (69 species), *Collembola* (16 species), and *Daphnia* (37 species).
3. A training alignment of 1623 pseudo-COI sequences and 150 species provided by the Data Analysis Working Group of the Consortium for the Barcode of Life (http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges/), from which five subtrees (each 20 to 57 species) were taken.

**RESULTS**

**Comparison of Simulated and Empirical Trees**

Sequence evolution was simulated on gene trees representing a broad range of phylogenetic situations (Fig. 1a). When tree height and intraspecific distance are plotted (Fig. 1b), the distribution of gene trees only partially matched the distribution of parameter values under which they were simulated. In the upper right quadrant of the log-log plot, representing shallow gene trees and deep coalescences, the simulated gene trees hit a threshold corresponding to situations where the tree depth equaled the coalescence depth. In these gene trees, the coalescences were effectively at the root of the tree. Otherwise, the computed parameters for each set of simulated trees matches reasonably well with the experimental scheme.

Trees estimated for empirical, or nearly so, mtDNA sequence alignments all had computed parameter values in the lower right quadrant of Figure 1, corresponding to theta \( \leq 0.1 \) and species birth rates \( \geq 10 \). For gene trees estimated from observed, not simulated, sequences, we can expect the depth to increase with the number of species represented, until the true root node has been included (Sanderson, 1996). In contrast, the number of sequences will affect only the variance of our estimate of intraspecific distance. Given that some of these empirical trees contain fewer than the 50 species of the simulated trees, the points representing empirical trees may be biased rightward. Overall, the lower right-hand quadrant of Figure 1a can be taken to represent gene trees estimated from mtDNA. Mitochondrial genes have an effective population size of one quarter that of nuclear genes. Consequently, trees estimated for nuclear genes might be expected to fall higher on these figures than the corresponding mtDNA trees.

**Species Monophyly**

The proportion of species that was monophyletic on each tree varied along the upward (positively sloped) diagonal axis from deep trees with shallow coalescences (species birth rate = 0.3, theta = 0.01), where it was nearly universal, to shallow trees with deep coalescences (species birth rate = 30, theta = 1), where it was nearly nonexistent (Fig. 2a). The sigmoidal manner of the variation in monophyly along this dimension was apparent both when the dimension was visualized as the product of theta and the species birth rate (Fig. 3a) and when it was visualized as the ratio of the intraspecific to interspecific (nearest heterospecific) distances (Fig. 3b). This ratio is a measure of species differentiation, with low values representing relatively large distances between species and high values representing less differentiation. The rate of monophyly did not vary on the orthogonal axis, from shallow trees with shallow coalescences to deep trees with deep coalescences.

To assess the effect of the number of sequences (gene copies sampled per species) on the incidence of monophyly, we reestimated it for each sample size, summed across replicates. We found a noticeable dependency between monophyly and sample size in those evolutionary scenarios that fell on or just above the downward (negatively sloped) diagonal of Figure 1a (Fig. 55; available online at www.systematicbiology.org). In these cases, monophyly decreased with increasing sample size until a species was represented by five to six sequences, above which the incidence of monophyly was constant. However, in other evolutionary scenarios, either below the downward diagonal of Figure 1a, or well above it, there was no effect of sample size.
b) Species Identification by Method

<table>
<thead>
<tr>
<th>Correct ID (species present)</th>
<th>Uncertain ID (species present)</th>
<th>Incorrect ID (species present)</th>
<th>False Positive ID (species absent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liberal tree-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strict tree-based</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Species Identification

The rates of identification outcomes in the different evolutionary scenarios are summarized by method in Figure 2b. For all of the methods, when the true species was represented in the data set, the rate of correct identification declined along the upward diagonal from deep trees with shallow coalescences to shallow trees with deep coalescences. Three of the methods (BLAST, distance, and liberal tree-based) had higher rates of correct identifications than the fourth, strict tree-based method across all evolutionary scenarios but were prone to making more incorrect identifications. The difference was due to the strict tree-based method returning a large number of uncertain identifications, especially in shallow trees with deep coalescences.

There were substantial differences among the methods when the true species was not represented in the data set. The rate at which BLAST methods made unjustified false-positive identifications increased as trees shortened. The distance methods behaved similarly to BLAST but appeared also to be influenced by monophyly or other features of the upward diagonal axis (see below). The liberal tree-based method showed a near-uniform moderate rate of false-positive identification, whereas the strict tree-based method had significantly lower false-positive rates, especially in deeper trees with shallower coalescences.

Four methods were tested for proposing species identifications from the results of a BLAST search. Three of the methods (BLAST1, BLAST3, and BLAST4) had very similar rates of correct identification across the space of simulated trees, with BLAST1 having marginally better success (Fig. S1; available online at www.systematicbiology.org). The fourth method (BLAST2) was significantly less successful in almost every case. The BLAST2 method was also most likely to return an incorrect identification when the true species was represented in the reference data set. All of the methods were equally likely to return an uncertain identification, when the E-value cut-off was exceeded. When the true species was not represented in the reference data set, only two outcomes were possible, uncertainty or false-positive identification. Again, all methods were equally likely to return an uncertain identification so consequently all methods had the same rate of false-positive identifications. Overall, the BLAST1 method, which is to accept the top hit as the species identification, performed best. All subsequent discussion of BLAST methods will refer only to this method.

Two distance methods were tested, assigning to the query the identity either of the nearest species on average (distance (mean)) or of the nearest sequence (distance (nearest)). The distance (nearest) method performed as well or better than distance (mean) when the true species was represented in the reference data set (Fig. S2; available online at www.systematicbiology.org) and equally well when the true species was not included. Given the superior performance of the distance (nearest) method, all subsequent discussion of the distance method will refer only to this method.

The tree-based methods were tested with a threshold to avoid query misplacement due to long-branch attraction. The threshold had no effect on the outcomes when the true species was present in the reference data set. When the true species was absent, the threshold reduced the rate of false-positive identifications by small amounts (~1%) in only three scenarios when the strict method was used, but by large amounts (10% to 80%) in 16 scenarios when the liberal tree-based method was used (Figs. S3 and S4; available online at www.systematicbiology.org).

Monophyly and Identification

To assess the relationship between species monophyly and the performance of each of the identification methods, the results were combined across tree space and plotted as a running average for both true- and false-positive identifications (Fig. 4). For BLAST, distance, and liberal tree-based methods, the rate of true-positive identifications varied identically (see also Fig. 2), falling approximately linearly with monophyly, to about 20% monophyletic, beyond which they fell precipitously. The strict tree-based method had a lower overall rate of true-positive identification, and this rate fell with a steeper slope than it did for the other methods. In contrast, when the true species was not represented in the reference data set, there were qualitative and quantitative differences among the methods. The BLAST and distance methods had similar trends, with the rate of false positives rising rapidly as monophyly fell. The liberal tree-based method varied independently of monophyly, as seen earlier in Figure 2, in strong contrast to the strict tree-based method, for which it rose slowly to a maximum of approximately 25% as monophyly declined to about 50%.

We showed earlier that monophyly varied sigmoidally with the product of theta and the species birth rate (Fig. 3a) and with the ratio of intraspecific to interspecific distances (Fig. 3b). When the rate of correct identifications is plotted against the distance ratio (Fig. 3c), we find that all methods have near maximum performance on trees where the intraspecific distance was ≤10% of the distance between the members of one species and its nearest heterospecific. As the ratio increased, reducing the difference between intra- and interspecific distances, the rate of successful identification declined rapidly. A very different trend appears when we consider identifications made when the correct species is not in the reference data set (Fig. 3d). The rate at which false-positive identifications were made by the BLAST, distance, and
FIGURE 3. The relationship of the ratio of mean intraspecific distance to the mean distance to the nearest heterospecific to monophyly and the rates of successful identification. (a) The sigmoidal variation in monophyly across the tree space upward diagonal dimension (Fig. 2) defined by the product of theta and the species birth rate. (b) A similar sigmoidal relationship between monophyly and the ratio of intra- and interspecific distances. (c) The proportion of correct identifications made when the true species is present. (d) The proportion of incorrect identifications made when the true species was absent. The lines in (d) enclose results from the same method. In (a) and (b), numbers indicate multiple data points. In (c), there are many instances of overlapping data points.

FIGURE 4. The proportion of true- and false-positive identifications made by each method in relation to the proportion of species that are monophyletic on the simulated tree. Each line represents a running average for three values of proportion monophyletic species (0.02 increments). The solid lines represent true-positive identifications, made when the true species was represented in the reference data set. The dashed lines represent false-positive identifications, made when the true species was not represented.
liberal tree-based methods was largely unrelated to the amount of species differentiation indicated by the ratio. On the other hand, the rate of false positives returned by the strict tree-based method increased as differentiation declined, with near-zero rates only occurring when the ratio was \(\leq 2\%\) to 5\%. When the liberal tree-based method was used with a threshold, the rate of false-positive identifications fell dramatically when the ratio was \(\leq 5\%\), but at no time was it as low as that observed with the strict tree-based method.

**Overlap in Distance Distributions**

The degree of overlap in the distributions of intraspecific and interspecific distances, the “barcoding gap,” has been used as a predictor of the reliability of barcoding, with both optimistic (Barrett and Hebert, 2005; Chase et al., 2005; Hajibabaei et al., 2006; Hebert et al., 2004; Janzen et al., 2005) and pessimistic (Meyer and Paulay, 2005) assessments. Here we found, for the trees in Figure 1c, that the overlap of the distribution of distances from the query sequence to each of its conspecifics on the distribution of distances to heterospecifics was a poor predictor of identification success when the distance method was used (Fig. 5a). Success rates in excess of 90\% still occurred when the distributions overlapped by up to 60\% to 80\%. At higher rates of overlap, the variance in rate of correct identification increased dramatically. Overlap is a poor predictor because the distribution of intraspecific distances is an inflated estimate of the distance between a query sequence and the closest representative sequence of the potentially correct species. If paraphyly is common and lineage sorting incomplete, then many species will be represented by multiple single-species clades, interspersed with clades of sister species. In computing all intraspecific distances, one is combining distances to other members of the same clade with distances to conspecifics in other clades. If heterospecifics occur in intervening clades, at intermediate distances, then an overlap must occur. However, because distance-based identification is based on the nearest sequence to the query, the barcoding gap should be assessed using distances to only the nearest conspecific and heterospecific sequences. These distributions will differ from those for all distances by truncating them downwards. However, we found that this measure of overlap remained a poor predictor of identification success (Fig. 5b). Although the mean rate of success declined with increased overlap, the variance in success increased correspondingly. Nevertheless, when the overlap was \(\leq 10\%\), success identification occurred with consistently high rates (98\% to 100\%). From this analysis, we may conclude that although there is a general trend for distance-based identification success to decline with distance overlap, the distance overlap is a poor predictor of identification success.

**Heterogeneity in Rates of Molecular Evolution among Lineages**

The effect of among-lineage rate heterogeneity was examined by selecting the species trees previously simulated under three evolutionary scenarios and then resimulating coalescent gene trees and sequence evolution using three values of the rate heterogeneity \(a\) parameter (Table S1; available online at www.systematicbiology.org). The differences in the rates of identification due to
variation in rate heterogeneity were miniscule relative to the differences among methods, speciation rates, or coalescence depths. There is also no hint of a correlation between the rates of true- and false-positive identifications and rate heterogeneity.

**Fixed versus Variable Theta**

In most of the analyses, theta was held constant across the tree. The effect of allowing theta to vary among species was tested for three evolutionary scenarios (Table S2; available online at www.systematicbiology.org). These combinations correspond to a "transect" along the gradient in species monophyly, in the lower right quadrant of Figure 1a. In each replicate, the value of theta varied over a threefold range. There were only slight changes in the rates of identification outcomes, usually much less than 1%, when theta varied among species. Whether a greater variation in theta may have had a greater effect on the rate of successful identification requires further study.

**Sensitivity to Threshold Values**

When the E-value exceeds a cut-off (BLAST) or when the genetic distance exceeds a threshold (distance and tree-based), the identification is ambiguous. The sensitivity of these methods to the values of these thresholds was tested by reidentifying the query sequences in all of the simulated data sets using different threshold values.

When a more restrictive \(10^{-6}\) BLAST E-value cut-off was compared with a more liberal \(10^{-2}\) value, the proportion of true-positive identifications fell by values in the range 0 to 0.04, uncertain identifications rose by a similar amount, and incorrect identifications fell by 0 to 0.02, across all data sets where the true species was represented. When the true species was absent, the more restrictive cut-off resulted in a reduction in the rate of false-positive identifications in the range 0.001 to 0.12 (Fig. 6a), with most of the change occurring at tree depths where there had been a rapid change in false-positive identification rates (Fig. 2b). The more restrictive cut-off effectively shifted the transition rightwards.

A more restrictive distance threshold was obtained by using the 90th percentile of intraspecific distance, as compared with the 99th percentile. In evolutionary scenarios where the coalescences were not at or near the root of the tree (theta \(\leq 0.3\)), the 90th percentile threshold was approximately one-half the value of the 99th percentile. When this more restrictive threshold was used, the proportion of true-positive identifications fell very slightly (range 0 to 0.01), uncertain identifications rose by a similar amount, and incorrect identifications did not change. In marked contrast, when the true species was absent, the more restrictive threshold reduced the false-positive rate by up to 0.3 over large regions of the tree space (Fig. 6b).

**FIGURE 6.** The change in the frequency of false-positive identifications when the threshold has a more liberal value instead of a more restrictive value, for (a) BLAST and (b) the distance method. The plots have the same organization as in Figure 2, but note the different scale.

**FIGURE 7.** The proportion of correct identifications of dipteran flies by the strict tree-based method calculated from Table 4 (revised identification criteria) in Meier et al. (2006), after removing singletons, plotted against the estimated ratio of intraspecific distance to distance to nearest heterospecific. Each data point corresponds to a genus and is coded by the mean number of reference sequences per species (Nseq/Sp). The lines summarize the results (Fig. 2b) obtained here by simulation.
Sensitivity to Number of Sequences

We reestimated the frequency of identification outcomes for each sample size, summed across replicates, for each evolutionary scenario (Fig. 1a). For both the BLAST and distance methods (Figs. S6 and S7; available online at www.systematicbiology.org), the frequency of correct identifications rose in a curvilinear fashion with the number of reference sequences for those scenarios where species differentiation was low (above the downward diagonal in Fig. 1a) but was nearly constant otherwise. A similar curvilinear relationship was observed for the liberal tree-based method (Fig. S8; available online at www.systematicbiology.org), but for those cases where a correct identification was not made, there appeared to be a fixed proportion of ambiguous, rather than incorrect, identifications. When the strict tree-based method was used (Fig. S9; available online at www.systematicbiology.org), the curvilinear relationship was much stronger, where in all cases the rate of correct identification was reduced for species represented by fewer than six sequences. This sample size dependency is clearly demonstrated in Figure 7, where traces of identification success converge as sequence numbers increase.

Discussion

No single method of species identification was superior across the range of phylogenetic scenarios studied (Fig. 2). When all species were represented in the reference data set, the BLAST, distance, and liberal tree-based methods were equally good, and better than the more conservative strict tree-based method. However, if the query species was not in the reference data set, then only the strict tree-based method could be relied upon not to make a falsely positive identification. In this situation, both the BLAST and distance methods gave high rates of identification in many different evolutionary scenarios when ambiguity was the correct outcome, and the liberal tree-based method was uniformly unreliable. Using a threshold in conjunction with the liberal tree-based method reduced the rate of false-positive identifications in some situations when the true species was absent, but this approach consistently made more false-positive identifications than did the strict tree-based method.

Under what evolutionary circumstances can we expect to obtain reliable identifications by the methods described here? Our results indicate that there is a clear requirement for species to be in well-differentiated clades. High rates of correct identification occurred when there was only slight (≤ 10%) overlap in the distributions of distances to nearest conspecifics and heterospecifics (Fig. 5), but greater overlap did not necessarily predict poorer success. Further, success occurred when the mean distance to a conspecific was only a small fraction (≤ 10%) of the mean distance to the nearest heterospecific (Fig. 3). Species monophyly per se has only a relatively slight effect, with success exceeding 90% when only 80% of species were monophyletic (Fig. 4).

The tree-based statistics that we inferred for a small number of empirical mtDNA data sets placed them in one quadrant of the phylogenetic scenarios that we simulated (Fig. 1c). The selection of empirical data sets was intended to be indicative, not exhaustive, and we make the conjecture that naturally occurring phylogenetic situations populate a wider tree space. There are clear patterns across tree space in the reliability of the identification methods tested (Figs. 2 and 3), and perhaps these patterns could only be identified by investigating extreme scenarios. These patterns are quite robust to simulated variation in rate heterogeneity among branches, model of sequence evolution, and effective population size. The empirical trees exhibit a wide range of species differentiation, with variation in the ratio of intraspecific to nearest heterospecific distances spanning the whole range illustrated in Figure 3. Consequently, it is not possible to predict whether these methods would be reliable with any specific group of organisms unless the relative levels of intra- and interspecific differentiation were known.

BLAST

The informal identification of an unknown specimen by using BLAST to search large public databases may be as reliable a method as any other. Further, applying a decision criterion that sums the “weight of evidence,” which integrates across top-ranking BLAST hits, is no more reliable than simply using the best hit. However, the reliability of BLAST is mainly dependent on the comprehensiveness of the taxon representation in the database. A major concern regarding the use of BLAST against the NCBI reference database is that the taxon sampling of this database reflects diverse research priorities and many taxa are unrepresented. This does not preclude the use of BLAST against a curated database containing all relevant species. Altering the E-value cut-off to more or less restrictive values will tune down or up the probability of BLAST incorrectly making a positive identification. The E-value, the probability of a random match having the observed quality, is proportional to the size of the sequence search space so that increasing either the number or length of sequences in the reference database will reduce the E-value of a given match. Consequently, this property makes it unreasonable to set a particular cut-off for application across databases and its choice is entirely pragmatic. One strategy for choosing a suitable cut-off might be to BLAST sequences from species in groups where taxon sampling has been thorough and then select a value that is less than the E-value associated with the highest-scoring heterospecific hits. The E-value cut-offs used here were relatively arbitrary as we were interested in studying variation across evolutionary scenarios and different parameter values.

Distance Methods

DNA barcoding was initially presented as a hybrid system based on a simple measure of sequence similarity and a rapid technique (NJ) of tree visualization (Hebert et al., 2003). More recently, as sequence data have accumulated, the emphasis has shifted to sequence
comparison, with tree-based analyses used as confirmation (Hajibabaei et al., 2006). Now DNA barcoding is explicitly based on distance comparisons (Ratnasingham and Hebert, 2007).

Meyer and Pauley’s (2005) assessment of DNA barcoding was the first to involve a well-characterized taxonomic group. Although it was performed using placement in an NJ tree as the test criterion, because each species was represented by a single exemplar sequence, it was effectively equivalent to an assessment based solely on sequence similarity. The method correctly identified 80% of the traditional species and 98% of phylogenetically defined evolutionary significant units (ESUs). They conclude that barcoding holds promise directly identified 80% of the traditional species and 98% of the phylogenetic scenarios sampled. A species might be unrepresented because it is either unsampled or unrecognized. The probability of this erroneous placement would be increased if the query is closely related to the members of the clade, as happens with cryptic species; more distantly related, but unrecognized, species are more likely to be placed deeper in the tree. A second situation is if there is significant paraphyly among the reference species. If an identical sequence occurred in two species, but was represented by only one species in the reference set, then the query might be misidentified. So, inadequate sampling within a species in combination with paraphyly could result in the placement of the query with the wrong species.

The strict tree-based method was significantly more conservative than the liberal approach in making a true-positive identification. This is not surprising, for if the true species is present in the reference data set, then by chance alone the query will fall sister to a single-species clade with a probability of $P = \frac{1}{n}$, where $n$ is the proportion of reference sequences in single-species clades containing $n$ sequences. Given the number of reference sequences used here (uniform distributions 1 to 19), we expect that the strict tree-based method will increase the rate of false-negative or ambiguous cases by approximately 13%, even when all species are monophyletic, and by much more when they are paraphyletic or polyphyletic. In the region of sampled tree space where monophyly was high (lower left quadrant, Fig. 2), the observed difference between the two tree-based methods was $\sim$10%, and as the rate of monophyly declined, so the difference in rate of false negatives increased, to a maximum of 60%. These rates clearly show the magnitude of the cost in terms of false negatives imposed by the criterion that the query may not fall sister to the true reference sequences.

The significance of the number of reference sequences is illustrated by the results of Meier et al. (2006), who tested distance and tree-based methods using dipteran sequences obtained from GenBank. They applied each method to each sequence, using the remaining sequences as the reference data set. Generally, low rates of successful identification were obtained but the potential outcomes were biased downwards because over 16% of species were represented by singletons. When the identity of these singletons was tested, they were removed from the reference alignment, leaving no representatives of the species in the references, and so could only ever give an ambiguous or incorrect result. Similarly, species having two sequences in their data set could only ever have one reference sequence when the other was being tested. All of the methods require at least one reference sequence per species so the dual use of singletons as query and reference sequences is to conflate issues of taxon sampling with method evaluation. For example, Meier et al.’s “best match” distance method gave about 68% success, but after correction for singletons this rises to 81%. Their tree-based method “revised identification
The critical role of the threshold is to limit false-positive, not incorrect, identifications. When first proposed by Hebert et al. (2004), a threshold of $10 \times$ the mean intraspecific distance was used to distinguish a conspecific from a heterospecific. Subsequently, Meyer and Paulay (2005) found that thresholds of $3.2 \times$ to $6.8 \times$ the mean intraspecific distance were optimal for minimizing error in the identification of marine gastropods. The 90th and 99th percentiles here used as thresholds with the distance method correspond, respectively, to $2.4 \times$ and $4.8 \times$ the mean intraspecific distance. Changing the threshold had almost no effect on the rate of correct or incorrect identifications, but relaxing the threshold significantly increased the rate of false-positive identifications when the species was not represented in the data set (Fig. 6b) in the lower right quadrant, in a region where empirical trees occurred (Fig. 1c) and where false positives covaried with monophyly (Fig. 2). Distance thresholds offer a solution to the problem for tree-based methods of an unknown being placed in a position sister to, or within, a single-species clade by chance when the true species is not in the reference data set. This technique has been used successfully to avoid the misidentification of noncetacean sequences as cetacean (Ross and Murugan, 2006).

The distance threshold was effective in limiting false-positive identifications only in those scenarios where species were very well differentiated (Fig. 3d). Both the distance and liberal tree-based methods made substantial numbers of false identifications when the distance to a conspecific was only 10% of the distance to nearest heterospecific. Although using the more stringent 90th percentile as the threshold (Fig. 6b) did reduce these false-positive rates, they still exceed those of the strict tree-based method. Using a threshold as great as $10 \times$ would be expected to give unacceptably high rates of false positives in most situations.

Hebert et al. (2004) also proposed that the $10 \times$ threshold could be used to define species boundaries and to aid in species discovery. Here we have observed that correct identification can be obtained reliably when no more than 10% of distances to nearest conspecifics overlap with heterospecifics, and when the ratio of distances to conspecifics to distances to nearest heterospecific is also $\leq 10\%$. However, unlike the scenario proposed by Hebert and colleagues, in our simulations all of the species were known and represented in the reference data set. This empirical 10% threshold (or “$10 \times$ rule” as termed by Hickerson et al., 2006) therefore represents the limit at which species can be identified reliably rather than a natural species boundary. Simulation studies by Hickerson et al., (2006) in fact indicate that the application of distance thresholds is unlikely to identify novel species of animals, primarily because the development of reproductive isolation generally is faster than the accumulation of differences in neutral marker loci such as COI. Perhaps this is the true “barcoding gap,” the distance between speciation and detection.

Thresholds

The critical role of the threshold is to limit false-positive, not incorrect, identifications. When first proposed by Hebert et al. (2004), a threshold of $10 \times$ the mean intraspecific distance was used to distinguish a conspecific from a heterospecific. Subsequently, Meyer and Paulay (2005) found that thresholds of $3.2 \times$ to $6.8 \times$ the mean intraspecific distance were optimal for minimizing error in the identification of marine gastropods. The 90th and 99th percentiles here used as thresholds with the distance method correspond, respectively, to $2.4 \times$ and $4.8 \times$ the mean intraspecific distance. Changing the threshold had almost no effect on the rate of correct or incorrect identifications, but relaxing the threshold significantly increased the rate of false-positive identifications when the species was not represented in the data set (Fig. 6b) in the lower right quadrant, in a region where empirical trees occurred (Fig. 1c) and where false positives covaried with monophyly (Fig. 2). Distance thresholds offer a solution to the problem for tree-based methods of an unknown being placed in a position sister to, or within, a single-species clade by chance when the true species is not in the reference data set. This technique has been used successfully to avoid the misidentification of noncetacean sequences as cetacean (Ross and Murugan, 2006).

The distance threshold was effective in limiting false-positive identifications only in those scenarios where species were very well differentiated (Fig. 3d). Both the distance and liberal tree-based methods made substantial numbers of false identifications when the distance to a conspecific was only 10% of the distance to nearest heterospecific. Although using the more stringent 90th percentile as the threshold (Fig. 6b) did reduce these false-positive rates, they still exceed those of the strict tree-based method. Using a threshold as great as $10 \times$ would be expected to give unacceptably high rates of false positives in most situations.

Hebert et al. (2004) also proposed that the $10 \times$ threshold could be used to define species boundaries and to aid in species discovery. Here we have observed that correct identification can be obtained reliably when no more than 10% of distances to nearest conspecifics overlap with heterospecifics, and when the ratio of distances to conspecifics to distances to nearest heterospecific is also $\leq 10\%$. However, unlike the scenario proposed by Hebert and colleagues, in our simulations all of the species were known and represented in the reference data set. This empirical 10% threshold (or “$10 \times$ rule” as termed by Hickerson et al., 2006) therefore represents the limit at which species can be identified reliably rather than a natural species boundary. Simulation studies by Hickerson et al., (2006) in fact indicate that the application of distance thresholds is unlikely to identify novel species of animals, primarily because the development of reproductive isolation generally is faster than the accumulation of differences in neutral marker loci such as COI. Perhaps this is the true “barcoding gap,” the distance between speciation and detection.

Paraphyly and Species Identification

Species-level paraphyly and polyphyly, known to be common among animals (Funk and Omland, 2003), are thought to jeopardize the reliability of genetic methods of identification. This is because of the possibility that the query sequence will have greatest similarity with, or be placed phylogenetically in the same clade as, a sequence from a related but incorrect species. One of the principal causes of species-level paraphyly, incomplete lineage sorting, is addressed by the simulations reported here and occurs when gene coalescences predate speciation events. Overall, this study indicates that paraphyly, although diminishing the accuracy of species identification, does so in a restricted fashion. Although it is possible that the simulated variation in theta, the coalescent depth, may have been less than that which occurs naturally, the methods tested here do not strictly require monophyly for a successful identification, as occurs with methods such as “Tree-based identification senso Hebert” (Meier et al., 2006). When species paraphyly occurs, identification errors are most likely to arise when species share haplotypes and lack differences rather than because of the paraphyly per se. So long as the sequences form single-species clades, the methods tested here will offer a certain degree of reliable identification.

Sampling Issues

Critical to the achievement of reliable identification when paraphyly occurs is the extent to which genetic variation has been sampled. If haplotypes shared by two species have not been sampled in both, then erroneous, rather than ambiguous, identifications will result. We found that when species were poorly differentiated,
all methods showed an improvement in success with increasing number of reference sequences. The strict tree-based method, which provided the best protection against making false-positive identifications when the species was unrepresented, also showed a method-specific dependency on sequence number. Our results indicate that species should be represented by five or more reference sequences to achieve best identification success. Such sampling may be more extensive than can be achieved for the majority of species, for whom singletons may be the norm. Whereas well-curated data sets of reference sequences have proven valuable, for those studying highly diverse and inconspicuous organisms (e.g., Ekrem et al., 2007), rather than large and charismatic species (Baker et al., 1996; Baker and Palumbi, 1994), such thorough sampling is nontrivial.

CONCLUSIONS

These simulations indicate that when all species are adequately represented in the reference data sets, genetic methods can give reliable species identifications. The degree to which species are genetically differentiated appears to be a critical determinant of success. When all species are represented in the reference data set, BLAST, distance, and liberal tree-based methods will be equally successful and make more correct identifications than the strict tree-based method, which requires that the query sequence must fall within, and not sister to, a single-species clade. The strict tree-based method is conservative, making ambiguous or false-negative identifications at a rate inversely proportional to the number of reference sequences per species.

When the correct species has not been included in the reference data set, only the tree-based methods, especially the strict method, coupled with a distance threshold will protect against false positives. The other methods are ubiquitously poor or have a rate of error determined by empirical thresholds.

One of the prime motivations for the development of genetic methods is their large-scale application to species identification. A major criticism of these methods has been that they will be unreliable because of inadequate sampling of genetic variation and incorrect taxonomy. These concerns can be mitigated by applying a conservative approach, by using the strict tree-based method. However, once the specific taxonomic group is well understood and its genetic diversity is fully sampled, this conservative approach is no longer warranted. It would be appropriate to switch to whichever of the other methods, BLAST, distance, or a more liberal tree-based approach, are the computationally most efficient and provide the greatest speed. The requirement for both well-differentiated species and multiple reference sequences per species, in order to achieve an acceptable level of successful identifications, may render these techniques inappropriate in some circumstances. In a finite world, there will always be a trade-off between the accuracy and the cost, measured in both time and money, of species identification. It is important that the reliabilities of different approaches are fully understood so that an informed decision may be made.

ACKNOWLEDGEMENTS

Assistance with computation was received from Matthew Goode, Thomas Lopdell, Marcel van de Steeg, and Pui Shan Wong. We thank Jack Sullivan, Marshal Hedin, and two anonymous reviewers for their helpful comments.

REFERENCES


First submitted 3 May 2007; reviews returned 13 July 2007; final acceptance 11 January 2008

Associate Editor: Marshal Hedin