MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 1848–1863 doi: 10.1111/j.1365-294X.2011.05235.x

A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding

A.-B. ZHANG,* C. MUSTER,† H.-B. LIANG,‡ C.-D. ZHU,‡ R. CROZIER,§[1] P. WAN,* J. FENG¶ and R. D. WARD**

*College of Life Sciences, Capital Normal University, Beijing 100048, China, †Zoological Institute and Museum, University of Greifswald, Greifswald, Germany, ‡Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China, §Department of Evolutionary Genetics, School of Marine and Tropical Biology, DB23, James Cook University, Townsville, Qld 4811, Australia, ¶College of Applied Mathematics, Capital Normal University, Beijing 100048, China, **Wealth from Oceans Flagship, CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart, Tas. 7001, Australia*

Abstract

Reliable assignment of an unknown query sequence to its correct species remains a methodological problem for the growing field of DNA barcoding. While great advances have been achieved recently, species identification from barcodes can still be unreliable if the relevant biodiversity has been insufficiently sampled. We here propose a new notion of species membership for DNA barcoding—fuzzy membership, based on fuzzy set theory—and illustrate its successful application to four real data sets (bats, fishes, butterflies and flies) with more than 5000 random simulations. Two of the data sets comprise especially dense species/population-level samples. In comparison with current DNA barcoding methods, the newly proposed minimum distance (MD) plus fuzzy set approach, and another computationally simple method, 'best close match', outperform two computationally sophisticated Bayesian and BootstrapNJ methods. The new method proposed here has great power in reducing false-positive species identification compared with other methods when conspecifics of the query are absent from the reference database.

Keywords: DNA barcoding, fuzzy set theory, species membership, statistical approach

Received 13 July 2010; revision received 27 April 2011; accepted 10 May 2011

Introduction

DNA barcoding (http://www.barcodinglife.org) has gained widespread prominence during the past 8 years as part of the worldwide campaign to develop a global biodiversity inventory (Hebert *et al.* 2003a,b; Ebach & Holdrege 2005; Gregory 2005; Marshall 2005; Savolainen *et al.* 2005; Schindel & Miller 2005; Ward *et al.* 2005; Abdo & Golding 2007; Hajibabaei *et al.* 2007a; Robin *et al.* 2007; Roe & Sperling 2007; Meusnier *et al.* 2008; Zhang *et al.* 2008; Monaghan *et al.* 2009; Ward *et al.* 2009; Dinca *et al.* 2011). On 11 April 2011, there were 1 181 714 barcode records from 99 138 species in the Barcode of Life Database (BOLD) (http://www.barcodinglife.org). Nonetheless, some reservations still remain about the utility of DNA barcoding (Moritz & Cicero 2004; Will & Rubinoff 2004; Prendini 2005; Brower 2006; Hickerson *et al.* 2006; Meier *et al.* 2006; Whitworth *et al.* 2007; Song *et al.* 2008; Silva-Brando *et al.* 2009; Lou & Golding 2010).

Species membership and its corresponding methodology have been among the most contentious and animated issues in the application of DNA barcode information to species identification and species circumscription (Hebert *et al.* 2003a,b; Hebert *et al.* 2004; Hickerson *et al.* 2006; Meier *et al.* 2006; Nielsen & Matz 2006; Rubinoff *et al.* 2006; Hajibabaei *et al.* 2007a,b; Munch *et al.* 2008a,b; Ross *et al.* 2008; Zhang *et al.* 2008; Chu *et al.* 2009; Lou & Golding 2010). Traditionally, most empirical DNA barcoding projects have applied classical phylogenetic approaches for species

Correspondence: Ai-Bing Zhang, Fax: +86 1068901860; E-mail: zhangab2008@mail.cnu.edu.cn

[1]Deceased.

assignments, such as neighbour-joining (Saitou & Nei 1987; Hebert *et al.* 2003a,b), maximum parsimony (Ekrem *et al.* 2007) and Bayesian methodology (Elias *et al.* 2007), or pure statistical approaches based on classification algorithms (Austerlitz *et al.* 2009). Some new methods have also been proposed, such as decision theory (Abdo & Golding 2007) and artificial intelligence-based approaches (Zhang *et al.* 2008).

Recently, considerable advances in species assignment via DNA barcoding have been achieved, especially through the framework of Bayesian theory (Munch *et al.* 2008a,b; Lou & Golding 2010). Bayesian methods provide the necessary statistical strength to distinguish between well- and poorly supported assignments and, most importantly, provide a measure of statistical confidence (Munch *et al.* 2008a,b; Lou & Golding 2010). Three Bayesian approaches have been proposed to date. The first is a method that calculates the likelihood of coalescence for sequences known to originate from a particular species and then calculates the change in likelihood when the query is considered a member of this species (Abdo & Golding 2007; Lou & Golding 2010). Coalescent methods can be time-consuming for large data sets owing to the huge number of coalescent trees generated to sample all possible coalescent events (Lou & Golding 2010). The second Bayesian method is the Statistic Assignment Package (SAP) that incorporates taxonomic information from NCBI and uses this information to impose topology constraints on the trees sampled from Markov Chain Monte Carlo (MCMC) algorithms. The probability of assignment is the number of sampled trees showing the likelihood of the query sequence branching with a sequence from a certain species (Munch *et al.* 2008a,b; Lou & Golding 2010). While the method in SAP was found to have good statistical performance on real and simulated data sets, it may not be easily applicable to large-scale data sets (Munch *et al.* 2008a,b). Therefore, Munch *et al.* (2008b) proposed a new Bayesian method using a constrained neighbour-joining method (hereafter referred to as BootstrapNJ) to accelerate the DNA assignments for large data sets and showed that the new method performs as well as the more computationally intensive full Bayesian approach (Munch *et al.* 2008b). The only drawback of this method is that it does not model species not present in the database and can lead to wrong inferences (Munch *et al.* 2008b). The third Bayesian method, recently proposed, is to accelerate the coalescent method (Abdo & Golding 2007) by replacing the coalescent-based MCMC algorithm with one that makes use only of the number of segregating sites from the sequences of a single species. A segregating sites method uses only sites at which there is a nucleotide change and therefore is very rapid. However, the method (Lou & Golding 2010) suffers from a loss of information by compressing the entire collection of sequence data into a single number, although the loss is assumed to be trivial. Despite some limitations, all these methods have greatly contributed to the success of the DNA barcoding campaign.

Apart from the complexity introduced by the aforementioned methodology, there have been other controversial debates on DNA barcoding issues, such as the threshold issue. Hebert *et al.* (2004) proposed the use of a divergence threshold (the '10 times rule'—10× the mean intraspecific variation for the group under study) to define species boundaries. The threshold approach proved to be useful in several groups of organisms, fishes (Ward *et al.* 2005), crustaceans (Lefebure *et al.* 2006), North American birds (Hebert *et al.* 2004), tropical lepidopterans (Hajibabaei *et al.* 2006) and cave-dwelling spiders (Paquin & Hedin 2004). The use of thresholds in species assignments has subsequently been extensively debated because of the lack of strong biological support and universality to all animal species (Meyer & Paulay 2005; Hickerson *et al.* 2006; Rubinoff *et al.* 2006; Ward *et al.* 2009). Meier *et al.* (2008) argue and document that barcoding gaps are often incorrectly computed. The use of mean instead of smallest interspecific distance exaggerates the size of the 'barcoding gap' and can lead to misidentification. A second issue relates to sampling. At the current stage of development of DNA barcoding reference databases, depth of intraspecific sampling is usually sacrificed in favour of greater taxonomic coverage (Matz & Nielsen 2005). For instance, a typical barcoding study includes only 5–10 individuals (sometimes only singletons) per species for the vast majority of species (Hajibabaei *et al.* 2007b; Zhang *et al.* 2010). Such sample sizes are unlikely to uncover all the genetic diversity of a population, let alone a species (Zhang *et al.* 2010). Besides insufficient intraspecific sampling, the database coverage of species sampling is also incomplete; currently, there is much undescribed species diversity (Rubinoff *et al.* 2006). Thus, for many, if not most, unknown specimens, established DNA barcoding databases do not yet permit the accurate assignment of species names. Most current methods will yield incorrect identifications for queries whose conspecifics are not present in the reference database. A third issue is monophyly (Farris 1974). Tree-based methods assume that species are monophyletic; this may be unrealistic, especially for recently diverged species (Hudson & Coyne 2002; Hickerson *et al.* 2006; Knowles & Carstens 2007; Lou & Golding 2010). For example, 17% of bird species deviated from mtDNA monophyly (Funk & Omland 2003), casting doubt on the precision of DNA barcoding for allocating individuals to previously described avian species. However, McKay & Zink (2010) found that a high

proportion of the reported paraphyly in Funk & Omland (2003) was because of poor taxonomy. Reciprocal monophyly at the species level was considered as a basic premise of correct identifications, especially for tree-based methods (Hudson & Coyne 2002; Hickerson *et al.* 2006; Nielsen & Matz 2006; Knowles & Carstens 2007). Nevertheless, barcoding is possible in the case of reciprocal paraphyly using combinations of mutations that are specific to a given species (Austerlitz *et al.* 2009). The fourth issue is gene sampling. Single gene-based barcoding (COI-based barcoding) was initially proposed by Hebert *et al.* (2003a,b) and has found great success in the animal groups mentioned above. Efficient COI-based barcoding has also been documented for a few groups of fungi (e.g. *Penicillium* sp., Seifert *et al.* 2007), macroalgae (Rhodophyta, Saunders 2005) and two ciliophoran protist genera (*Paramecium* and *Tetrahymenas*, Barth *et al.* 2006; Lynn & Struder-Kypke 2006; Chantangsi *et al.* 2007). However, COI barcoding does not resolve plant species, where the use of several plastid genes has been recommended (Newmaster *et al.* 2006; Ferri *et al.* 2009). Thus, it is now commonly accepted that in some groups, multiple-gene barcoding is required (Newmaster *et al.* 2006; Ferri *et al.* 2009), and *matK* and *rbcL* have been selected as plant barcode markers by CBoL (http://www.barcoding.si.edu/plant_working_group.html). While sometimes necessary, a multiple-locus system exacerbates problems with primer design, especially where primers are not universal across groups. It increases labour and consumable costs and increases the chance of having an incomplete reference data set.

These various issues largely reflect the problem of having incomplete information, whether it be insufficient representatives of each species in the reference set, incomplete database coverage of related species, a short segment of DNA instead of the whole genome and so on. However, even if this additional information were available, ecological, behavioural and other biological data relating to species identification have not been incorporated into most DNA barcoding studies. Barcode species allocations are generally only based on DNA information. In the situation of insufficient information concerning the species studied, there are two principal ways to deal with the data: one is the development of sophisticated statistical methods, such as the Bayesian methods mentioned above, and another is based on the fuzzy set method. The former has been extensively studied (Munch *et al.* 2008a,b; Abdo & Golding 2007; Lou & Golding 2010), while the potential of the latter has not yet been explored with respect to DNA barcoding. We here propose a fuzzy-theory-based species identification approach to analyse species membership in DNA barcoding.

Fuzzy sets were introduced by Zadeh (1965) as an extension of the classical notion of sets. In fuzzy sets, elements have degrees of membership. Unlike classical set theory, where the membership of an element in a set is assessed in binary terms, fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described with the aid of a membership function valued in the real unit interval [0, 1]. Fuzzy sets generalize classical sets, because the indicator functions of classical sets are special cases of the membership functions of fuzzy sets, taking only values 0 or 1. Fuzzy set theory can be used in a wide range of domains with incomplete or imprecise information, such as bioinformatics (Liang *et al.* 2006).

We suggest that, based only on a single or a few genes, and without ecological, behavioural or any other biological information, species identifications via DNA barcoding cannot be totally determined, but that assignations can be made as a fuzzy member of a named species. We demonstrate our method on four real data sets using a combination of a minimum distance (MD) method (Ross *et al.* 2008) with the fuzzy membership function (FMF) values proposed in this study. We compare our approach to several currently using methods [Bayesian method, BootstrapNJ, 'best close match' (BCM)], using more than 5000 random replications in total.

Materials and methods

Fuzzy membership

Definition of membership function for a species. Assume that there be an unknown query sequence q which may or may not belong to a species A according to its genetic distance (or any other criterion) from A. The query sequence is taken to be of the potential same species as that of the reference sequence with the smallest pairwise distance from the query sequence. Furthermore, x is defined as the genetic distance between a query and the known species A. Based on fuzzy set theory, q is called not included in a fuzzy set A if $f(x) = 0$, q is called fully included if $f(x) = 1$ and q is called a fuzzy member if $0 < f(x) < 1$. Mathematically, a fuzzy set is a pair (A,f) where A is a set and $f{:}A \to [0,1]$. For each $x \in A$, $f(x)$ is called the grade of x in (A,f), the species membership function $f(x)$ (Fig. 1). This function is defined in detail by the following equation (Lin *et al.* 2005; Yuan *et al.* 2008):

$$f(x;\theta) = \begin{cases} 1, & x \le \theta_1 \\ 1 - 2\left(\frac{x-\theta_1}{\theta_2-\theta_1}\right)^2, & \theta_1 \le x \le \frac{\theta_1+\theta_2}{2} \\ 2\left(\frac{x-\theta_1}{\theta_2-\theta_1}\right)^2, & \frac{\theta_1+\theta_2}{2} \le x \le \theta_2 \\ 0, & x \ge \theta_2 \end{cases} \quad (1)$$

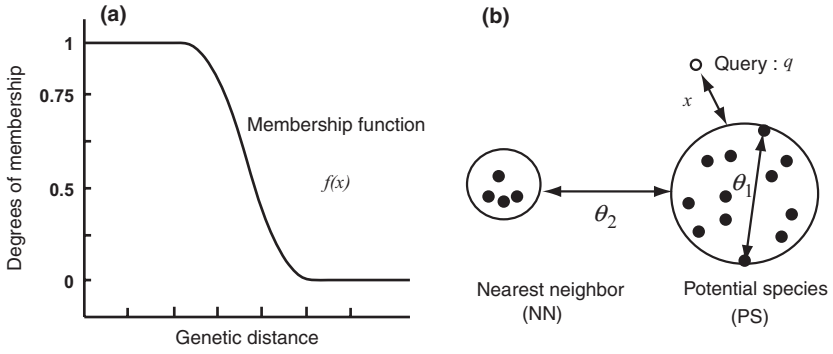

Fig. 1 Curve of membership function $f(x)$ and the estimation of two parameters, θ_1 and θ_2. (a) Curve of membership function $f(x)$; (b) estimation of θ_1 and θ_2. Black dots indicate DNA sequences for individuals in that species, while the empty dot means each query sequence. x is measured as genetic distance, and K2P distance (Kimura 1980) is used in this study.

Two parameters θ_1 and θ_2 need to be estimated based on the actual data set.

Estimation of parameters for membership function f(x). To determine the parameters θ_1 and θ_2 of eqn 1, which are maximum intraspecific and minimum interspecific genetic distances, respectively, the method searches for the nearest neighbour (NN) of that potential species (PS, Fig. 1b) of a query. Once the NN is found, the minimum K2P genetic distance of that species to the PS estimates the interspecific variation and the maximum K2P distance among the individuals of the PS estimates the intraspecific variation, to cover as much genetic variation within species as possible. K2P distance (Kimura 1980) was chosen as it has widespread use in the barcoding literature. However, potential nonreciprocal monophyly of the PS and its NN may result in a small or zero 'barcoding gap'. In this case, we instead use the fifth percentile of all interspecific variation and the 95th percentile of all species in the database to represent genetic variation between and within species, respectively.

After determining the parameters of eqn 1, the FMF values can be easily obtained. To calculate an FMF value, first the K2P distance of the query to the PS identified by the MD method mentioned above or any other method has to be computed, and then, the resultant distance value is input to eqn 1. To test the robustness of FMF values on the success rate of species identification, we used three different fuzzy function values (FMF = 0.90, 0.95 and 0.99; denoted as MF90, MF95 and MF99) as thresholds for accepting (if higher) or rejecting (if lower) species membership.

Comparison to the existing methods

A general strategy to make comparisons among different methods. We wished to determine whether our new method performs better than current DNA barcoding methods, including Bayesian (Munch *et al.* 2008a), BootstrapNJ (Munch *et al.* 2008b) and BCM (Meier *et al.* 2006) methods, utilizing 'leave-one-out' simulation. In these simulations, we remove one sequence at a time and use it as a query, with all other sequences remaining as the reference database. We performed 500 random replications for each empirical data set and each method except for the two methods implemented in the computer program SAP (Munch *et al.* 2008a,b). For these, we performed 100 random replications to save computation time. Sequences belonging to species present only once (referred to as 'singletons' hereafter) were especially important as queries to test the efficiency of the fuzzy approach. This is because these species will no longer be represented in the reference data set, and most current methods will therefore make misidentifications. For the remaining sequences, we examined whether those approaches work well when the reference data set does have a conspecific sequence.

Algorithms and the computer program packages. The Bayesian approach is based on a combination of automated database searches, alignments and Bayesian phylogenetic inferences whose objective is to approximate the posterior probability that the unknown specimen belongs to a particular species or taxonomic group (Munch *et al.* 2008a). This MCMC-based approach is computationally demanding. The BootstrapNJ method (Munch *et al.* 2008b) uses a neighbour-joining algorithm (Saitou & Nei 1987) in combination with bootstrapping (Felsenstein 1985) as a heuristic approach to approximate the posterior probabilities (Munch *et al.* 2008b). The Bayesian and BootstrapNJ methods used here are implemented in the program package SAP (Munch *et al.* 2008a,b). SAP version 1.08 and the latest version 1.12 were downloaded and installed locally on a linux system. An in-house database constructed from each empirical data set was annotated using the taxonomic information from NCBI or assembled manually.

The SAP program has 47 options that can be set by users. These include six general options, eight for NetBlast search, 21 for homologue set compilation, one for Alignment, four for Output and others. It is very

difficult to test them all. However, they all have been assigned default values for good reasons (Munch *et al.* 2008a,b), and we used the default values in all our comparative simulations.

The BCM identification protocol first identifies the best barcode match of a query but only assigns the species name of that barcode to the query if the barcode is sufficiently similar. This approach requires a threshold similarity value that defines how similar a barcode match needs to be before it can be identified. Such a value could be estimated for a given data set by obtaining a frequency distribution of all intraspecific pairwise distances and determining the threshold distance below which 95% of all intraspecific distances are found. The BCM approach is implemented in the computer program TaxonDNA (Meier *et al.* 2006).

The MD method and the calculation of the FMF value are implemented in a new program package FuzzyIdentification which was developed in the current study (available at http://life.cnu.edu.cn/soft/FuzzyIdentification.rar). The original program was written in the Matlab language and has been compiled into a windows executable file (.exe) for users, obviating the need to install Matlab itself.

Success rate of species identification and confidence intervals. The success rate of species identification is defined by the following formula (Zhang *et al.* 2008):

$$\mathrm{Rate}_{\mathrm{success}} = \frac{\mathrm{Number}_{\mathrm{hit}}}{\mathrm{Number}_{\mathrm{test}}} \tag{2}$$

where $\mathrm{Number}_{\mathrm{hit}}$ and $\mathrm{Number}_{\mathrm{test}}$ are the numbers of sequences successfully hit by methods under study and the number of total query sequences examined, respectively. A successful hit is counted as such if a query is assigned to its correct species name in the database or a potential misidentification is flagged by low FMF values.

For the MD plus fuzzy set approach, a successful hit is further defined when (i) the MD method makes a correct sequence assignment and the fuzzy set approach generates a high FMF values (three different thresholds, MF90, MF95 and MF99, were used); or (ii) the MD method makes a wrong assignment, but the fuzzy set approach produces a low FMF value (below the threshold values used), especially for singletons.

Binary data indicating the presence (successful identification) or absence (failed identification) of a specific attribute are often modelled as random samples from a Bernoulli distribution with parameter prob, where prob is the proportion in the population with that attribute. A $(1 - \alpha)$-level confidence interval (CI) for prob is calculated by the following formula (Tamhane & Dunlop 2000):

$$\frac{(\widehat{\mathrm{prob}} - \beta)}{(1 + \frac{z^2}{n})} \leq \mathrm{prob} \leq \frac{(\widehat{\mathrm{prob}} + \beta)}{(1 + \frac{z^2}{n})} \tag{3}$$

where $\alpha = 0.05$, $\beta = \sqrt{[\widehat{\mathrm{prob}}(1 - \widehat{\mathrm{prob}})z^2]/n + z^4/4n^2}$ and $z = z_{\alpha/2}$ (n is the number of replications and z is the critical value corresponding to an area $1 - \alpha$ under the standard normal curve).

Empirical data sets and their analysis

To evaluate this new notion of species membership—fuzzy membership, as defined here—we used four empirical data sets downloaded from the BOLD system (http://www.barcodinglife.org), representing different scales of data sets and genetic diversities. As identification difficulties arise when the unknown specimens come from a currently underdescribed part of biodiversity (Rubinoff 2006; Rubinoff *et al.* 2006), we especially examined cases in which the conspecifics of queries were not represented in the reference set of these data sets (bats, fishes, butterflies and flies) by using singletons as queries. We used three 'clean' data sets (ambiguous sites, such as 'Ns', removed: bats, fishes and butterflies) and one 'raw' data set (ambiguous sites and gaps kept: flies). The reason for using clean data sets (ideal cases) is to facilitate fair comparisons among the different DNA barcoding methods, as different methods may treat ambiguous sites in different ways. On the other hand, the 'raw' data set serves as a more practical case where ambiguous sites, gaps and slightly different sequence lengths are kept.

To test whether the method suggested here can identify the correct species or correctly report false-positive identifications when the conspecifics of queries are absent from the reference set, we examined queries from both nonsingletons and singletons. We applied the MD method, and the MD method in conjunction with the FMFs method, to identify the queries against the complete reference set (for nonsingleton queries) and against the incomplete one (singleton queries), respectively.

Two sorts of tests were subsequently conducted to calculate (i) the total or overall success rate of species identification (including singletons and nonsingletons), (ii) the success rate for singletons and (iii) the success rate for nonsingletons. In the first test, a sequence was randomly removed from the data set and used as the query, regardless of singleton status, with remaining sequences as reference sequences. This was repeated 500 times. This enabled the total success rate for species identification and nonsingleton species identification to be calculated following eqn 2, with their corresponding 95% CI computed using eqn 3. In the second test, each

singleton was chosen as the query, the remaining sequences forming the reference library. This was repeated many times depending on the number of singletons in the database. The MD method will always assign a species name from the database to the query sequence, although this may be wrong. If the fuzzy approach reports a low fuzzy membership value for the singleton query, this is counted as a success (a low fuzzy membership value indicates a potential misidentification). The success rate of species identification for singletons and its 95% CI were calculated as above following eqns 2 and 3.

The bats and fishes data sets The COI data set of 87 Neotropical species of bats in Guyana contained 840 COI sequences with a length >600 bp, from 47 genera (Clare *et al.* 2007). We cleaned the data set by removing sequences with ambiguous sites, such as 'Ns', and those whose length were <648 bp, the standard length in COI DNA barcoding (Hebert *et al.* 2003a,b; Hebert *et al.* 2004). Sequence alignment used the program MUSCLE (Edgar 2004) with the default setting to check sequence homology; 766 COI sequences representing 84 bat species remained for the subsequent analysis (Table 1).

Steinke *et al.* (2009) barcoded 201 North Pacific fish species, yielding 1225 barcode sequences. We downloaded these sequences from the BOLD (http://www.barcodinglife.org) project Fishes of Pacific Canada Part I. Read lengths were about 655 bp long. We removed uncertain nucleotide sites, such as 'Ns'. The resultant 652 bp alignments of 982 sequences and 188 species (Table 1) were used in the subsequent analysis.

The butterflies data set A complete DNA barcode data set for a country's butterfly fauna was recently reported (Dinca *et al.* 2011). This comprised the 180 species of butterflies in Romania (some one-third of the European butterfly fauna). Morphology and DNA barcodes of more than 1300 specimens belonging to six families (Hesperiidae, Papilionidae, Pieridae, Lycaeridae, Riodinidae and Nymphalidae) were studied. We downloaded sequences from GenBank (accession numbers HQ003941 to HQ005268) (Dinca *et al.* 2011). The data set was cleaned by removing sequences with ambiguous sites, such as 'Ns', and those whose lengths were <648 bp; 1235 sequences from 174 species remained (Table 1). Sequence alignment used the program MUSCLE (Edgar 2004) with the default setting.

The flies data set Meyer and Paulay's data set of 263 taxa of cowries is considered as one of the most dense species- and population-level barcode data sets (Meyer & Paulay 2005). However, we were unable to obtain the full data set of that study by downloading it from their webpage (http://www.flmnh.ufl.edu/cowries), by accessing the GenBank numbers they provided or by contacting the authors. Meier *et al.* (2006) published a comprehensive fly study. We instead downloaded this data set that included a high proportion of singletons (71.71%) among the 449 dipteran species (Table 1). This provided us with an excellent opportunity to test our fuzzy set approach and the other methods. Unlike our previous treatments, we did not clean the data set, retaining the original alignment of Meier *et al.* (2006) including all ambiguous sites and indels ('gaps').

Results and discussions

Bats and fishes

After data cleaning, we obtained 766 COI sequences representing 84 bat species to test all methods under study (Table 1). In the case of the total query, the MD plus fuzzy set approach achieved a 98.2% success rate

Table 1 The four empirical data sets used in this study and fuzzy membership function values (FMFs) for true and false identifications over more than 5000 replications

Taxa group	No. sequences	No. species (singletons)	No. query replicatinos	FMF for true-positive identification	FMF for false-positive identification	Sensitivity/ specificity
Bats*	766†	84 (9)	1200 (36‡)	0.98	0.12	0.99/0.88
Fishes	982	188 (45)	1200 (180)	0.99	0.26	0.96/0.96
Butterflies	1235	174 (9)	1200 (36)	0.99	0.34	0.96/0.89
Flies	1333	449 (321)	1000 (642)	0.95	0.08	0.82/1.00

*Bats and fishes data sets were taken from the webpage of the Barcode of Life Database System http://www.barcodinglife.org/; temperate Europe butterfly data set was downloaded from Dinca *et al.*'s study (2011) and the fly data of Diptera were retrieved from these studies (Meyer & Paulay 2005; Meier *et al.* 2006).

†Clean size of the data, but the fly data (Meyer & Paulay 2005; Meier *et al.* 2006), in which ambiguous sites of DNA sequences and the original alignment were kept to test the robustness of the methods being compared with each other here, see also text.

‡The number of replications from singletons for all methods used.

of species identification with 95% CI from 96.61% to 99.05% over the 500 random queries (Fig. 2, Table 2, Appendix S2, Supporting information). The BCM method gave a 1.27% similarity threshold to define species boundaries. Both the BCM and MD methods obtained a 98.4% success rate of species identification with 95% CI from 96.87% to 99.19% (Fig. 2, Table 2, Appendix S2, Supporting information). These results suggest that the newly proposed MD plus fuzzy set approach performs equally as well as the extant BCM method because there are no statistically significant differences in the success rate of species identification between them. However, both the new method and the BCM method outperform significantly both Bayesian and BootstrapNJ methods. These methods correctly identified only 25 and 59 queries, respectively, over 100 random queries with 25% (95% CI: 17.55–34.3%) and 59% (95% CI: 49.20–68.13%) success rates.

There are only nine singletons in the bat data set. Bayesian and BootstrapNJ methods failed to make species assignments at the species level for the nine queries using these singletons. The BCM method would lead to a misidentification of all nine, for example, the query of BCBN27905 from species '*Furipterus horrens*' was assigned to species name '*Platyrrhinus helleri*' (Appendix S2, Supporting information). The MD method alone also assigned all these singleton queries to incorrect species names (Appendix S2, Supporting information). However, eight of these nine identifications were assigned very low FMF values, ranging from 0.00 to 0.25, indicating that these identifications were very unlikely to be correct. For instance, the singleton query BCBN27905 of species '*F. horrens*' was identified as '*P. helleri*' using the MD method, but with an FMF value of 0.00, indicating a likely false identification (Appendix S2, Supporting information). One singleton query, BCBN31805, had

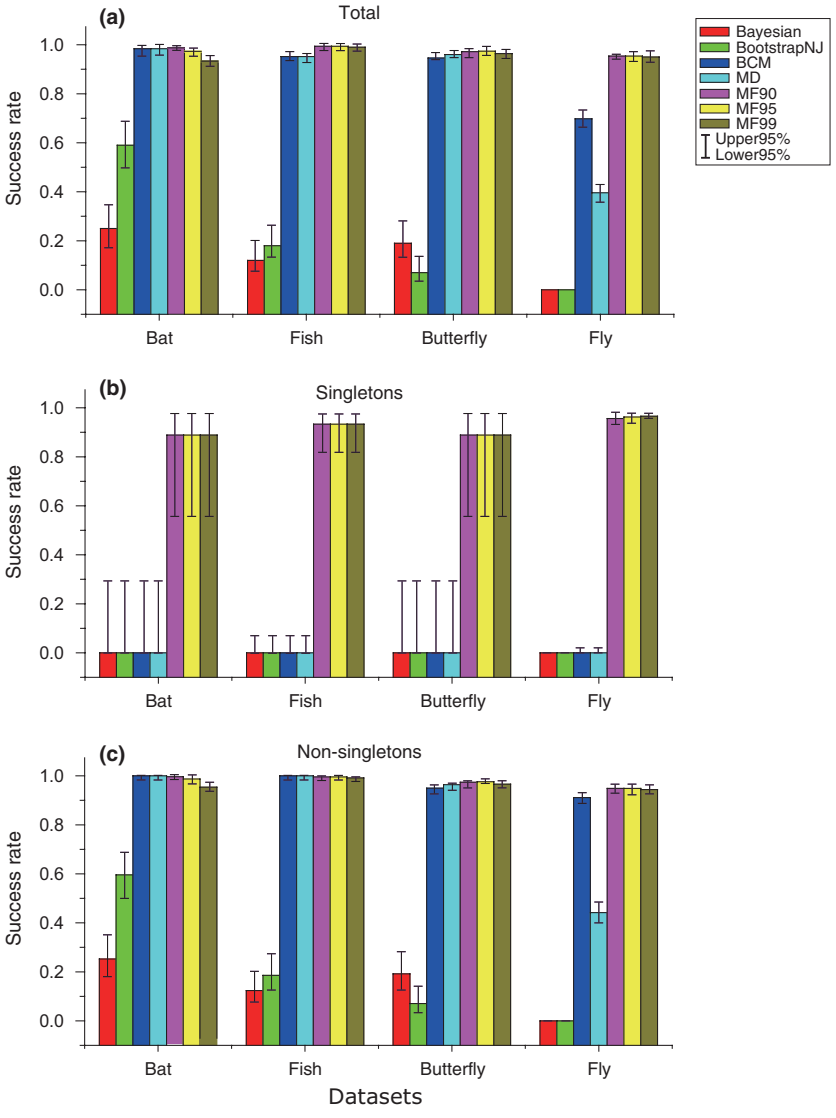

Fig. 2 Success rate of species identification and corresponding 95% confidence intervals for total queries from both singletons and nonsingletons, from singletons and from nonsingletons with leave-one-out simulations. More than 5000 simulations were performed based on four empirical data sets with different barcoding methods [Bayesian, BootstrapNJ, best close match, minimum distance (MD) and the MD plus fuzzy set approach]. (a) Total success rate of species identifications from both singleton and nonsingleton queries; (b) success rate of species identifications from singleton queries; (c) success rate of species identifications from nonsingleton queries.

Table 2 Species assignments for Neotropical bats based on COI sequences (Clare *et al.* 2007) through a minimum distance (MD) method together with fuzzy membership function values (FMFs) for 500 random replication queries and all singleton queries with leave-one-out simulations

Query types*	No.	Query†	Species assigned‡	Status§	FMF value¶
500 random replication queries	1	BCBN15205-Carollia perspicillata	*Carollia perspicillata*	(✓)	1.00
	2	BCBNT56506-Saccopteryx bilineata	*Saccopteryx bilineata*	(✓)	1.00
	3	BCBN77205-Lophostoma schulzi	*Lophostoma schulzi*	(✓)	1.00
	4	BCBNT32506-Sturnira tildae	*Sturnira tildae*	(✓)	1.00
	5	BCBN10905-Carollia perspicillata	*Carollia perspicillata*	(✓)	1.00
	6	BCBNT08306-Carollia brevicauda	*Carollia brevicauda*	(✓)	1.00
	7	BCBNC02706-Vampyrum spectrum	*Vampyrum spectrum*	(✓)	1.00
	8	BCBNC10706-Carollia perspicillata	*Carollia perspicillata*	(✓)	0.98
	9	BCBN77005-Phylloderma stenops	*Phylloderma stenops*	(✓)	1.00
	10	BCBNC00206-Myotis riparius	*Myotis riparius*	(✓)	1.00
	11	BCBNT12606-Choeroniscus minor	*Choeroniscus minor*	(✓)	1.00
	12	BCBN53605-Lonchophylla thomasi	*Lonchophylla thomasi*	(✓)	1.00
	13	BCBNT04106-Cormura brevirostris	*Cormura brevirostris*	(✓)	1.00
	14	BCBNT90806-Artibeus obscurus	*Artibeus obscurus*	(✓)	1.00
	15	BCBNT95006-Lonchophylla thomasi	*Lonchophylla thomasi*	(✓)	1.00
	16	BCBN14105-Lonchorhina inusitata	*Lonchorhina inusitata*	(✓)	1.00
	17	BCBNT62806-Trachops cirrhosus	*Trachops cirrhosus*	(✓)	1.00
	18	BCBNT23706-Vampyressa thyone	*Vampyressa thyone*	(✓)	1.00
	...	...	...	...	...
	500	BCBNT14306-Diclidurus isabellus	*Diclidurus isabellus*	(×)	1.00
Singleton queries	1	BCBN27905-Furipterus horrens	*Platyrrhinus helleri*	(×)	0.00
	2	BCBN52305-Saccopteryx gymnura	*Saccopteryx bilineata*	(×)	0.00
	3	BCBNT57206-Cyttarops alecto	*Rhynchonycteris naso*	(×)	0.00
	4	BCBN31805-Eptesicus chiriquinus	*Eptesicus furinalis*	(×)	0.93
	5	BCBNT39206-Lampronycteris brachyotis	*Micronycteris hirsuta*	(×)	0.00
	6	BCBNT11206-Lasiurus atratus	*Molossus* sp.	(×)	0.00
	7	BCBNT29006-Nyctinomops macrotis	*Molossus rufus*	(×)	0.00
	8	BCBNT38806-Molossus sp.	*Molossus molossus*	(×)	0.00
	9	BCBN12005-Cynomops planirostris	*Cynomops paranus*	(×)	0.25

*Five hundred random replication queries were performed with leave-one-out simulation in the first test, all singletons were taken out each to test the reliability of the method in the second test, see also text.

†The names of query sequences consist of accession numbers and their true species names after the dash, the table here presented only part of the results (19/500), see Appendix S2 (Supporting information) for all the assignments.

‡Species names are assigned by the MD method and FMF values, see also text.

§Ticks and crosses indicate correct and wrong assignments, respectively, by the MD method.

¶FMF value for single query.

a high FMF value of 0.93. This specimen had originally been identified as '*Eptesicus chiriguinus*' but our approach identifies it as '*Eptesicus furinalis*'. Its high FMF value suggests either that these two species are extremely close genetically or a need to recheck the original identification of this specimen. It is clear that the MD plus fuzzy set approach outperforms the other methods tested here in avoiding potential false species identification.

In the case of nonsingleton queries, where conspecifics were present in the reference database, the BCM, MD and MD plus fuzzy set approaches all achieved high success rates of species identification from 98.17% to 100%, while Bayesian and BootstrapNJ methods obtained low success rates (Bayesian, 25.25%, with 95% CI of 17.73–34.62%; BootstrapNJ, 59.6%, with 95% CI of 49.75–68.73%).

The fish data set comprised a 652-bp alignment of 982 COI sequences for 188 fish species (Table 1). The data set consisted of 937 nonsingletons and 45 singletons. In the 100 random simulations, Bayesian and BootstrapNJ methods identified just 12 and 18 nonsingleton queries to their correct species giving only 12.37% and 18.57% success rates (95% CI: 7.22–20.39% and 12.08–27.44%) (Fig. 2, Table 3, Appendix S3, Supporting information). In the 500 random queries simulation, 492 nonsingletons remained. These queries were successfully assigned to their correct species with a 100% suc-

Table 3 Species assignments for Pacific Canadian marine fish (Steinke *et al*. 2009) based on COI sequences through a minimum distance (MD) method together with fuzzy membership function values (FMFs) for 500 random replication queries and all singleton queries with leave-one-out simulations

Query types*	No.	Query†	Species assigned‡	Status§	FMF value¶
500 random replications queries	1	TZFPB61806-Anoplopoma fimbria	*Anoplopoma fimbria*	(✓)	1.00
	2	TZFPA12006-Paraliparis sp.	*Paraliparis* sp.	(✓)	1.00
	3	TZFPB10505-Eopsetta jordani	*Eopsetta jordani*	(✓)	1.00
	4	TZFPA03506-Paraliparis paucidens	*Paraliparis paucidens*	(✓)	1.00
	5	TZFPA19207-Nectoliparis pelagicus	*Nectoliparis pelagicus*	(✓)	1.00
	6	TZFPB58406-Citharichthys sordidus	*Citharichthys sordidus*	(✓)	1.00
	7	TZFPA14907-Eptatretus stoutii	*Eptatretus stoutii*	(✓)	1.00
	8	TZFPA19107-Oncorhynchus tshawytscha	*Oncorhynchus tshawytscha*	(✓)	1.00
	9	TZFPB47006-Cryptacanthodes aleutensis	*Cryptacanthodes aleutensis*	(✓)	1.00
	10	TZFPB46806-Porichthys notatus	*Porichthys notatus*	(✓)	1.00
	11	TZFPB77806-Theragra chalcogramma	*Theragra chalcogramma*	(✓)	1.00
	12	TZFPB09605-Rhacochilus vacca	*Rhacochilus vacca*	(✓)	1.00
	13	TZFPB44405-Dasycottus setiger	*Dasycottus setiger*	(✓)	1.00
	14	TZFPB58106-Allosmerus elongatus	*Allosmerus elongatus*	(✓)	1.00
	15	TZFPB33605-Bothrocara molle	*Bothrocara molle*	(✓)	1.00
	16	TZFPB63906-Bathyagonus pentacanthus	*Bathyagonus pentacanthus*	(✓)	1.00
	17	TZFPB16405-Psettichthys melanostictus	*Psettichthys melanostictus*	(✓)	1.00
	18	TZFPB15005-Alosa sapidissima	*Alosa sapidissima*	(✓)	1.00
	...	...	...	...	...
	500	TZFPB80806-Bathyagonus pentacanthus	*Bathyagonus pentacanthus*	(✓)	1.00
Singletons queries	1	TZFPB39005-Somniosus pacificus	*Squalus acanthias*	(×)	0.35
	2	TZFPB42705-Cyclothone pacifica	*Cyclothone atraria*	(×)	0.70
	3	TZFPB35905-Bathyraja spinicauda	*Bathyraja abyssicola*	(×)	0.11
	4	TZFPB16105-Raja binoculata	*Raja rhina*	(×)	0.09
	5	TZFPA08406-Syngnathus leptorhynchus	*Rimicola muscarum*	(×)	0.00
	6	TZFPA20907-Rimicola muscarum	*Anotopterus nikparini*	(×)	0.00
	7	TZFPB20005-Pleuronichthys coenosus	*Pleuronichthys decurrens*	(×)	0.09
	8	TZFPB19105-Lepidopsetta polyxystra	*Lepidopsetta bilineata*	(×)	0.31
	9	TZFPA17607-Chilara taylori	*Oligocottus maculosus*	(×)	0.00
	...	...	...	...	...
	45	TZFPA18007-Psychrolutes sigalutes	*Psychrolutes phrictus*	(×)	0.00

*Five hundred random replication queries were performed with leave-one-out simulation in the first test, all singletons were taken out each to test the reliability of the method in the second test, see also text.

†The names of query sequences consist of accession numbers and their true species names after the dash, the table here presented only part of the results (19/500), see Appendix S3 (Supporting information) for all the assignments.

‡Species names are assigned by the MD method and FMF values, see also text.

§Ticks and crosses indicate correct and wrong assignments, respectively, by the MD method.

¶FMF value for single query.

cess rate for both BCM and MD methods(95% CI: 99.20–100%) and a 99.57% success rate for the MD plus fuzzy set approach method (95% CI: 98.47–99.88% for MF95).

For the 45 singleton queries, the MD plus fuzzy set approach outperforms all the methods compared here, again by assigning low FMF values to queries to avoid false-positive identifications. For example, the singleton query TZFP39005 '*Somniosus pacificus*' was identified as species '*Squalus acanthias*' but with a low FMF value of 0.35 indicating a likely misidentification (Appendix S3, Supporting information), and the query TZFPA20907 '*Rimicola muscarum*' was assigned to '*Anotopterus nikaparini*' with an FMF value of 0.00. Bayesian and BootstrapNJ methods failed to assign these singletons at the species level. The BCM method assigned these singleton queries to their BCM species in the database, even though their true conspecifics were necessarily unrepresented in the database.

In the case of total queries (singletons and nonsingletons included), the MD plus fuzzy set approach achieved the highest success rate of species identification (99.4% with 95% CI: 98.25–99.8%) of all the methods under study (Appendix S3, Supporting information). Bayesian and BootstrapNJ obtained extremely low success rates of species identification (12% and 18%, respectively). The BCM and MD methods obtained relatively high success rates of species identification (95% with 95% CI of 92.96–96.75% for both), but still significantly less than that of the MD plus fuzzy set approach.

Butterflies and flies

The DNA barcode data set of the complete butterfly fauna for a whole country (Dinca *et al.* 2011) provided another good example to test the performance of the newly proposed method against several currently used methods. After clean-up, the butterfly data set comprised 1235 COI sequences of 174 butterfly species, and only nine singletons were found (Table 1). In the first test with 500 random replications, 498 nonsingletons were randomly chosen as queries each against the corresponding reference library. The MD plus fuzzy set approach achieved the highest species identification success rate of 97.59% (95% CI of 95.84–98.62%), while the MD method alone obtained a 96.38% success rate (Fig. 2, Table 4, Appendix S4, Supporting information). The BCM method gave a slightly lower success rate (94.98% with 95% CI of 92.69–96.58%). Both Bayesian and BootstrapNJ methods showed very low success rates (Bayesian, 19.2% with 95% CI of 12.65–28.05%; BootstrapNJ 7.07% with 95% CI of 3.47–13.88%). For the nine singletons, which were consequently not represented in the database, other approaches could not flag these as a likely misidentification. However, the MD plus fuzzy set approach identified eight of these nine as potential false-positive identifications by their low FMF values of 0.00 to 0.16 (Appendix S4, Supporting information). The only exception is gi|304270751. Here, *Hipparchia volgensis* was identified as *Hipparchia semele* with an FMF value of 1.00: these two species are recognized to share barcodes (Dinca *et al.* 2011). With total queries, the MD plus fuzzy set method outperforms all other methods and has the highest success rate of species identification (97.4% with 95% CI of 95.6–98.47%). The BCM and MD methods obtained slightly lower success rates, from 94.6% to 96.0% (Appendix S4, Supporting information), while both Bayesian and BootstrapNJ method obtained extremely low success rates (19% and 7% respectively).

The fly data set (Meier *et al.* 2006) comprises 1333 COI sequences and 449 species and was reported as a difficult data set for DNA barcoding with a relatively low species identification success rate. This data set serves as a good test case for DNA barcoding methods as it has two special features: dense species-level sampling and a high proportion of singletons (71.71% or 321 of the 449 species) (Table 1). The original alignment with all ambiguous sites of the data set was kept to examine all the methods under study with an uncleansed data set. Both the Bayesian and BootstrapNJ methods, as implemented in the program SAP (Munch *et al.* 2008a,b), failed to make any sequence assignments as they crashed (two different versions of SAP, 1.08 and the latest version of 1.12, were tested).

Of the 321 singleton queries, 309 potential false-positive identifications by the MD method could be recognized and avoided by their low FMF values with a mean of 0.08 (Tables 1 and 5, Fig. 2, Appendix S5, Supporting information), while all other methods failed to flag any of these as misidentifications.

In the 500 random replication simulation, the MD plus fuzzy set approach achieved a 94.86% success rate of species identification (95% CI: 92.41–96.55%) for the randomly selected 448 nonsingleton queries, whereas the MD method alone obtained a considerably lower success rate (44.19% with 95% CI of 39.66–48.82%). The BCM method showed a relatively high success rate for nonsingletons (91.12% with 95% CI of 87.85–93.58%), but still slightly less than that of the MD plus fuzzy set approach. In the situation of mixed queries taking both singletons and nonsingletons into account, the BCM method obtained a 69.8% success rate of species identification (with 95% CI of 65.64–73.66%), which is consistent with the success rate previously reported (Meier *et al.* 2006). The MD plus fuzzy set approach achieved an appreciably higher success rate, 95.4% (with 95% CI of 93.19–

Table 4 Species assignments for temperate Europe butterfly (Dinca *et al.* 2011) based on COI sequences through a minimum distance (MD) method together with fuzzy membership function values (FMFs) for 500 random replication queries and all singleton queries with leave-one-out simulations

Query types*	No.	Query†	Species assigned‡	Status§	FMF value¶
500 random replications queries	1	gi|304270609-Euphydryas aurinia	*Euphydryas aurinia*	(✔)	1.00
	2	gi|304271787-Pontia edusa	*Pontia edusa*	(✔)	1.00
	3	gi|304270889-Leptotes pirithous	*Leptotes pirithous*	(✔)	1.00
	4	gi|304269613-Apatura metis	*Apatura metis*	(✔)	1.00
	5	gi|304269985-Carcharodus alceae	*Carcharodus alceae*	(✔)	1.00
	6	gi|304271013-Lycaena thersamon	*Lycaena thersamon*	(✔)	1.00
	7	gi|304270871-Leptidea sinapis	*Leptidea sinapis*	(✔)	1.00
	8	gi|304271653-Plebejus sephirus	*Plebejus sephirus*	(✔)	1.00
	9	gi|304271079-Maculinea arion	*Maculinea arion*	(✔)	1.00
	10	gi|304270869-Leptidea sinapis	*Leptidea sinapis*	(✔)	1.00
	11	gi|304271639-Lycaeides argyrognomon	*Lycaeides argyrognomon*	(✔)	1.00
	12	gi|304271739-Polyommatus dorylas	*Polyommatus dorylas*	(✔)	1.00
	13	gi|304271565-Pieris rapae	*Pieris rapae*	(✔)	1.00
	14	gi|304270497-Erebia melas	*Erebia melas*	(✔)	1.00
	15	gi|304271107-Maculinea teleius	*Maculinea teleius*	(✔)	1.00
	16	gi|304271879-Pyrgus armoricanus	*Pyrgus armoricanus*	(✔)	1.00
	17	gi|304270701-Hipparchia fagi	*Hipparchia syriaca*	(×)	0.00
	18	gi|304271217-Melitaea cinxia	*Melitaea cinxia*	(✔)	1.00
	...	...	...	...	...
	500	gi|304271515-Pieris mannii	*Pieris mannii*	(✔)	1.00
Singletons queries	1	gi|304270731-Hipparchia statilinus	*Arethusana arethusa*	(×)	0.00
	2	gi|304270751-Hipparchia volgensis	*Hipparchia semele*	(×)	1.00
	3	gi|304270445-Erebia gorge	*Erebia epiphron*	(×)	0.00
	4	gi|304270917-Limenitis reducta	*Limenitis populi*	(×)	0.01
	5	gi|304270915-Limenitis populi	*Limenitis reducta*	(×)	0.01
	6	gi|304271377-Nymphalis l	*Nymphalis xanthomelas*	(×)	0.00
	7	gi|304269563-Allancastria cerisyi	*Zerynthia polyxena*	(×)	0.16
	8	gi|304271849-Pyrgus andromedae	*Pyrgus sidae*	(×)	0.00
	9	gi|304271675-Polyommatus amandus	*Polyommatus thersites*	(×)	0.00

*Five hundred random replication queries were performed with leave-one-out simulation in the first test, all singletons were taken out each to test the reliability of the method in the second test, see also text.
†The names of query sequences consist of accession numbers and their true species names after the dash, the table here presented only part of the results (19/500), see Appendix S4 (Supporting information) for all the assignments.
‡Species names are assigned by the MD method and FMF values, see also text.
§Ticks and crosses indicate correct and wrong assignments, respectively, by the MD method.
¶FMF value for single query.

96.92%), but the MD method alone obtained a very low success rate (39.6%, with 95% CI of 35.41–43.95%).

Processing time

The data analyses in this study were performed on a Red Hat Enterprise Linux Server (release 5.1, Tikanga, CPU:Intel Xeon CPU E5410 @ 2.33 GHz ×8) for SAP (Munch *et al.* 2008a,b) and on a 3.00-GHz desktop computer [Intel(R) Core (TM)2, DuoCPU, E8400 @ 3.00 GHz ×2] for BCM (Meier *et al.* 2006) and the MD plus fuzzy set approach. Hence, it is not possible to provide true comparative results for these methods performed on the two quite different systems. However, it is still useful for users to have approximate computation times needed to run these algorithms. SAP spent 2–8 min per assignment on the linux system, depending on data set size and the detailed algorithms of the Bayesian and BootstrapNJ methods, while the MD plus fuzzy set approach spent 3–6 min per assignment on the windows system, depending on the data set size (from 765 to 1332 reference sequences in this study). The strategy we used for BCM meant that it was much faster (a few seconds per assignment). BCM was performed once for all pairwise queries and a perl script used to calculate the success rate of species identification over the 500

Table 5 Species assignments for 449 species of Diptera (Meyer & Paulay 2005; Meier *et al*. 2006) based on COI sequences through a minimum distance (MD) method together with fuzzy membership function values (FMFs) for 500 random replication queries and all singleton queries with leave-one-out simulations

Query types*	No.	Query†	Species assigned‡	Status§	FMF value¶
500 random replications queries	1	gi|18032899-Lycoriella mali	*Lycoriella mali*	(✔)	1.00
	2	gi|25990016-Drosophila recens	*Drosophila recens*	(✔)	1.00
	3	gi|29029473-Culicoides imicola	*Culicoides imicola*	(✔)	1.00
	4	gi|28071168-Asphondylia yushimai	*Asphondylia yushimai*	(✔)	1.00
	5	gi|25989970-Drosophila subquinaria	*Drosophila subquinaria*	(✔)	0.94
	6	gi|11993743-Chiastocheta macropyga	*Anopheles gambiae*	(×)	0.00
	7	gi|13429940-Drosophila bocki	*Drosophila kikkawai*	(×)	0.40
	8	gi|29373386-Paragus politus	*Musca domestica*	(×)	0.00
	9	gi|18032871-Lycoriella mali	*Lycoriella mali*	(✔)	1.00
	10	gi|9799520-Phytomyza glabricola	*Phytomyza glabricola*	(✔)	1.00
	11	gi|26190182-Sapromyza mauli	*Anopheles gambiae*	(×)	0.00
	12	gi|21727859-Lucilia sericata	*Lucilia sericata*	(✔)	1.00
	13	gi|6979501-Acerocnema macrocera	*Drosophila paulistorum*	(×)	0.00
	14	gi|15724409-Aedes aegypti	*Aedes aegypti*	(✔)	1.00
	15	gi|21215164-Anopheles dunhami	*Anopheles stephensi*	(×)	0.00
	16	gi|7263046-Apocephalus paraponerae	*Anopheles gambiae*	(×)	0.00
	17	gi|14335215-Aedes punctor	*Anopheles gambiae*	(×)	0.00
	18	gi|8132648-Chironomus tenuistylus	*Drosophila emarginata*	(×)	0.00
	...	...	...	...	...
	500	gi|29409286-Lucilia sericata	*Aedes aegypti*	(×)	0.00
Singletons queries	1	gi|6979497-Acanthocnema glaucescens	*Drosophila tropicalis*	(×)	0.00
	2	gi|6979501-Acerocnema macrocera	*Drosophila paulistorum*	(×)	0.00
	3	gi|24430571-Aedeomyia squamipennis	*Sarcophaga cooleyi*	(×)	0.00
	4	gi|14335187-Aedes cantans	*Aedes pullatus*	(×)	0.00
	5	gi|14335193-Aedes cataphylla	*Aedes punctor*	(×)	0.86
	6	gi|14335195-Aedes cinereus	*Drosophila bipectinata*	(×)	0.00
	7	gi|14335203-Aedes geniculatus	*Drosophila azteca*	(×)	0.00
	8	gi|14335209-Aedes pullatus	*Aedes cataphylla*	(×)	0.00
	9	gi|14335217-Aedes rusticus	*Drosophila nebulosa*	(×)	0.00
	10	gi|14335223-Aedes vexans	*Apocephalus paraponerae*	(×)	0.00
	11	gi|30267329-Agathomyia unicolor	*Drosophila lusaltans*	(×)	0.00
	12	gi|30267379-Alipumilio avispas	*Drosophila saltans*	(×)	0.00
	...	...	...	...	...
	321	gi|29420568-Xylota ignava	*Calliphora livida*	(×)	0.00

*Five hundred random replication queries were performed with leave-one-out simulation in the first test, all singletons were taken out each to test the reliability of the method in the second test, see also text.
†The names of query sequences consist of accession numbers and their true species names after the dash, the table here presented only part of the results (19/500), see Appendix S5 (Supporting information) for all the assignments.
‡Species names are assigned by the MD method and FMF values, see also text.
§ticks and crosses indicate correct and wrong assignments, respectively, by the MD method.
¶FMF value for single query.

random queries. Other strategies will require BCM to spend much more time querying each sequence.

Conclusions

We propose a new notion of species membership—fuzzy membership—for use in DNA barcoding studies, where typically only DNA sequence information is used to identify species, and ecological, behavioural and other biological information is missing or incomplete. We applied this new concept to four real barcode data sets (bats, fishes, butterflies and flies) to solve their memberships and found it worked well. Barcode misidentifications arising from an incomplete reference data set of species could be flagged and recognized as misidentifications by low FMF values. The fuzziness of species membership proposed here is somewhat similar to the 'probability of assignment' values for query matches in the BOLD system, although they are different conceptually.

One may argue philosophically that an actual specimen or individual sampled could not in real life belong with an uncertain membership to a certain species (i.e. the fuzzy membership concept outlined here), because a species is a discrete evolutionary entity or unit. However, we may note that when only a short DNA sequence (partial COI for most animal species, perhaps several short plastid DNA segments for plants) is sampled from the entire genome, the sampled sequence cannot fully represent that specimen in a strict sense. Species boundaries can be easily blurred by homoplasy or interspecific hybridization (Zhang *et al.* 2005; Rubinoff *et al.* 2006). In some instances, different species share an identical barcode haplotype. Therefore, we argue that DNA barcode-based species identification can only be fuzzy. As has been claimed by Frezal & Leblois (2008), species boundary is not a definitive but a revisable dynamic concept (Rubinoff *et al.* 2006).

Database coverage will often be incomplete, as there will be much un-barcoded biodiversity, especially in the early stages of barcoding projects. We examined the notion of fuzzy membership of species in conjunction with a MD method in three different scenarios: mixed queries (total queries), singleton queries and nonsingleton queries. Our results from four real data sets show that for each of them, when the true species are included in the reference set, the MD method can assign unknown query sequences to their correct species with a high success rate of species identification. This is, of course, because these species usually have unique sequences—haplotypes are very rarely, in these instances, shared across species. Fuzzy function values for each identification of 1.00 (the theoretical upper limit) or approaching 1.00 further confirmed these correct identifications. These results are totally concordant with those from several other studies (Munch *et al.* 2008a,b; Ross *et al.* 2008). However, most methods return false-positive identifications when conspecifics of the queries are not represented in the database, with the NNs of the queries being improperly allocated (generally as congeners, if congeners are in the database). Our study on four real data sets demonstrate that a MD method in combination with fuzzy function values will greatly reduce the chance of accepting false-positive identifications by generating extremely low FMF values in such instances.

In principle, a complete reference database of all life is the key prerequisite for reliable species identification, and this is the goal of the DNA barcoding initiative (Hebert *et al.* 2003a,b; Hebert *et al.* 2004). The achievement of such a complete reference database for all life on the earth is still far from being realized and may be hindered or slowed by some methodological issues (such as nonuniversality of primers) and by biological factors such as NUMTs (Song *et al.* 2008). Biologists will face an incomplete barcoding database for a long time. However, information on distantly related species is arguably redundant in the database when identifying queries. Therefore, adopting a sequential investigation could reduce the total data set to the genus or family of the query sequence by first searching the complete database with a simple method, such as the genetic similarity method which has high speed (Rubinoff *et al.* 2006; Ross *et al.* 2008); then, much computational time can be saved with this much reduced and well-curated database to refine species identification with more reliable methods. However, for a given unknown query, one cannot know whether its conspecifics have been sampled in the reference set, and if they have not, then wrong assignments will be inevitable using most current methods (Munch *et al.* 2008b). Our fuzzy-theory-based approach proves to be a good alternative and, as we have shown on real data sets with more than 5000 random queries, can help to avoid false-positive identifications. We have presented three different fuzzy function values (MF90, MF95 and MF99) to test the robustness of FMF values on the success rate of species identification (see above and Appendix S1, Supporting information for details). They produced highly consistent results with each other (Fig. 2; Appendix S1a–c, Supporting information).

Comparisons to other methods were made. Our study indicates that computationally or logically complicated methods, such as the Bayesian and BootstrapNJ approaches, may not perform well in DNA barcoding. Our simulation with more than 5000 random queries for four empirical data sets shows that the computationally simple BCM method and the MD plus fuzzy set approach newly proposed in this study outperform both Bayesian and BootstrapNJ methods significantly in most situations. Species identification via DNA barcoding can be made without phylogenetic reconstruction; therefore, simple methods such as the BCM and the MD plus fuzzy set approach can make sequence assignments computationally easily. However, it is possible that the poor performance of Bayesian and BootstrapNJ methods seen herein relates to the default settings we used when running the program SAP, although 'The default settings of this program have been set with good reasons' (Munch *et al.* 2008a,b); the alternative of fine tuning the 47 options before running is not easy for most users.

In the case of queries from singletons, the newly proposed MD plus fuzzy set approach significantly outperforms all other methods under study. As mentioned above, a complete reference barcode library of all life is necessary for accurate identification of all queries, but this ultimate goal is very hard to reach. For some time,

most data sets will lack some species. Our tests using singleton queries necessitated depleting that species from the reference database, resulting in an incomplete data set and unrecognized misidentifications by most current methods. In such instances, the fuzzy set approach substantially reduced the risk of false-positive identification by generating low FMF values, effectively flagging such queries as likely misidentifications. Our method may in fact also be used in combination with any other method, such as a Bayesian approach, as an alternative means for elucidating species membership.

Acknowledgements

We are grateful to mathematician Prof W. Zhao, (College of Applied Mathematics, Capital Normal University, China) and Prof E. S. Tavares (Department of Natural History, Royal Ontario Museum, 100 Queen's Park, Toronto, Canada) for their kind help and useful comments on the manuscript. We also gratefully acknowledge the constructive comments of Dr Loren Rieseberg, Dr Francois Rousset and three anonymous referees on an earlier version of the manuscript. This study was supported by Beijing Municipal Natural Science Foundation Key Projects (grant no. KZ201010028028 to Zhang) and by Natural Science Foundation of China (to Zhang, grant no. 31071963, to Liang, grant no. 30570213), by Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (to Zhang, grant no. PHR201107120), by The Research Fund for the Doctoral Program of Higher Education (to Zhang, grant no. 20101108120002), by Grant from Public Welfare Project from the Ministry of Agriculture, China (grant no. 200803006), to Zhu. RHC's work on evolutionary genetics was supported by the Australian Research Council (grant no. DP0665890).

References

Abdo Z, Golding GB (2007) A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology*, **56**, 44–56.

Austerlitz F, David O, Schaeffer B *et al.* (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, **10**(Suppl 14), S10.

Barth D, Krenek S, Fokin SI *et al.* (2006) Intraspecific genetic variation in *Paramecium* revealed by mitochondrial cytochrome c oxidase 1 sequences. *The Journal of Eukaryotic Microbiology*, **53**, 20–25.

Brower AVZ (2006) Problems with DNA barcodes for species delimitation: 'ten species' of *Astraptes* fulgerator reassessed (Lepidoptera: Hesperiidae). *Systematics & Biodiversity*, **4**, 127–132.

Chantangsi C, Lynn DH, Brandl MT *et al.* (2007) Barcoding ciliates: a comprehensive study of 75 isolates of the genus *Tetrahymena*. *International Journal of Systematic & Evolutionary Microbiology* , **57**, 2412–2425.

Chu KH, Xu M, Li CP (2009) Rapid DNA barcoding analysis of large datasets using the composition vector method. *BMC Bioinformatics*, **10**(Suppl 14), S8.

ClareEl, Lim BK, Engstron MD *et al.* (2007) DNA barcoding of neotropical bats:Species identification and discovery within Guyana. *Molecular Ecology Notes*, **7**, 184–190.

Dinca V, Zakharov EV, Hebert PD *et al.* (2011) Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 347–355.

Ebach MC, Holdrege C (2005) DNA barcoding is no substitute for taxonomy. *Nature*, **434**, 697.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Ekrem T, Willassen E, Stur E (2007) A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics & Evolution*, **43**, 530–542.

Elias M, Hill RI, Willmott KR *et al.* (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 2881–2889.

Farris JS (1974) Formal definitions of paraphyly and polyphyly. *Systematic Zoology*, **4**, 548–554.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Ferri G, Alu M, Corradini B *et al.* (2009) Forensic botany: species identification of botanical trace evidence using a multigene barcoding approach. *International Journal of Legal Medicine*, **123**, 395–401.

Frezal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. *Infection, Genetics & Evolution*, **8**, 727–736.

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: Frequency, causes, consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology and Systematics*, **34**, 397–423.

Gregory TR (2005) DNA barcoding does not compete with taxonomy. *Nature*, **434**, 1067.

Hajibabaei M, Janzen DH, Burns JM *et al.* (2006) DNA barcodes distinguish species of tropical *Lepidoptera*. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 968–971.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007a) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.

Hajibabaei M, Singer GA, Clare EL *et al.* (2007b) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, **5**, 24.

Hebert PDN, Cywinska A, Ball SL *et al.* (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.

Hebert PDN, Ratnasingham S, deWaard JR *et al.* (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, **270**(Suppl), 96–99.

Hebert PDN, Penton EH, Burns JM *et al.* (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14812–14817.

Hickerson MJ, Meyer CP, Moritz C *et al.* (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, **55**, 729–739.

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution*, **56**, 1557–1565.

Kimura M (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, **56**, 887–895.

Lefebure T, Douady CJ, Gouy M *et al.* (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. *Molecular Phylogenetics & Evolution*, **40**, 435–447.

Liang LR, Lu SY, Wang XN *et al.* (2006) FM-test: A fuzzy-set-theory-based approach to differential gene expression data analysis. *BMC Bioinformatics*, **7**(Suppl 4), S7.

Lin LZ, Qian ZQ, Zong G *et al.* (2005) Study of computer self-learning method for the membership function coefficient in fuzzy diagnosis. *Research & Discussion*, **10**, 57–58.

Lou M, Golding GB (2010) Assigning sequences to species in the absence of large interspecific differences. *Molecular Phylogenetics and Evolution*, **56**, 187–194.

Lynn DH Struder-Kypke MC (2006) Species of Tetrahymena identical by small subunit rRNA gene sequences are discriminated by mitochondrial cytochrome c oxidase I gene sequences. *The Journal of Eukaryotic Microbiology*, **53**, 385–387.

Marshall E (2005) Taxonomy – will DNA bar codes breathe life into classification. *Science*, **307**, 1037.

Matz MV, Nielsen R (2005) A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1969–1974.

McKay BD, Zink RM (2010) The causes of mitochondrial DNA gene tree paraphyly in birds. *Molecular Phylogenetics & Evolution*, **54**, 647–650.

Meier R, Shiyang K, Vaidya G *et al.* (2006) DNA barcoding and taxonomy in *Diptera*: A tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Meier R, Zhang G, Ali F *et al.* (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. *Systematic Biology*, **57**, 809–813.

Meusnier I, Singer GA, Landry JF *et al.* (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, **9**, 214.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, 2229–2238.

Monaghan MT, Wild R, Elliot M *et al.* (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology*, **58**, 298–311.

Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*, **2**, 279–354.

Munch K, Boomsma W, Huelsenbeck JP *et al.* (2008a) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, **57**, 750–757.

Munch K, Boomsma W, Willerslev E *et al.* (2008b) Fast hylogenetic DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **363**, 3997–4002.

Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. *Canadian Journal of Botany*, **84**, 335–441.

Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. *Systematic Biology*, **55**, 162–169.

Paquin P, Hedin M (2004) The power and perils of ''molecular taxonomy'': a case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. *Molecular Ecology*, **13**, 3239–3255.

Prendini L (2005) Comment on 'Identifying spiders through DNA barcoding'. *Canadian Journal of Zoology*, **83**, 498–504.

Robin VV, Freek TB, Joop JA (2007) DNA barcoding reveals hidden species diversity in Cymothoe (Nymphalidae). *Proceedings of the Netherlands Entomological Society Meeting*, **18**, 95–103.

Roe AD, Sperling FAH (2007) Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics & Evolution*, **44**, 325–345.

Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, **57**, 216–230.

Rubinoff D (2006) Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology*, **20**, 1026–1033.

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for ''barcoding'' identification. *Journal of Heredity*, **97**, 581–594.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology & Evolution*, **4**, 406–425.

Saunders, GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1879–1888.

Savolainen V, Cowan RS, Vogler AP *et al.* (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1805–1811.

Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature*, **435**, 17.

Seifert KA, Samson RA, Dewaard JR *et al.* (2007) Prospects for fungus identification using C01 DNA barcodes with *Penicillium* as a test case. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 3901–3906.

Silva-Brando KL, Lyra ML, Freitas AV (2009) Barcoding Lepidoptera: current situation and perspectives on the usefulness of a contentious technique. *Neotropical Entomology*, **38**, 441–451.

Song H, Buhay JE, Whiting MF *et al.* (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 13486–13491.

Steinke D, Zemlak TS, Boutillier JA *et al.* (2009) DNA barcoding of Pacific Canada's *Marine Biology*, **156**, 2641–2647.

Tamhane AC, Dunlop DD (2000) *Statistics and Data Analysis: from Elementary to Intermediate*, 1st edn, pp. 288. Published by Pearson Education Asia Ltd. and Higher Education Press, Beijing.

Ward RD, Zemlak TS, Innes BH *et al.* (2005) DNA barcoding Australia's fish species. *Proceedings of the Royal Society B: Biological Sciences*, **360**, 1847–1857.

Ward RD, Hanner R, Hebert PDN (2009) The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, **74**, 329–356.

Whitworth TL, Dawson RD, Magalon H *et al.* (2007) DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1731–1739.

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

Yuan J, Shi HB, Liu C (2008) Construction of fuzzy membership functions based on least squares fitting. *Control & Decision*, **23**, 1263–1271.

Zadeh LA (1965) Fuzzy sets. *Information & Control*, **8**, 338–353.

Zhang AB, Kubota K, Takami Y *et al.* (2005) Species status and phylogeography of two closely related *Coptolabrus* species (Coleoptera: Carabidae) in South Korea inferred from mitochondrial and nuclear genes. *Molecular Ecology*, **14**, 3823–3841.

Zhang AB, Sikes DS, Muster C *et al.* (2008) Inferring species membership using DNA sequences with back-propagation neural networks. *Systematic Biology*, **57**, 202–215.

Zhang AB, He LJ, Crozier RH *et al.* (2010) Estimation of sample sizes for DNA barcoding. *Molecular Phylogenetics & Evolution*, **54**, 1035–1039.

A-B.Z. is a researcher at Capital Normal University and studies DNA barcoding/DNA taxonomy for animals (Insects, Lepidoptera), is especially interested in theoretical aspects of DNA barcoding/DNA taxonomy. C.M., H-B.L., C-D.Z., and R.D.W. are interested in DNA barcoding for other animal groups (spiders, ground beetles, bees, and fishes). P.W. and J.F. are bioinformaticians.

Data accessibility

Final DNA sequence assembly: Data deposited at Dryad: 10.5061/dryad.9037

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Success rate of species identification and corresponding 95% confidence intervals for total queries from both singletons and nonsingletons, from singletons and from nonsingletons with leave-one-out simulations (detailed).

Appendix S2 (a–d) Detailed species assignments for Neotropical bats (Clare *et al.* 2006) based on COI sequences for 500 replications of queries [100 for Bayesian and NJ + Bootstrap (BootstrapNJ) methods] with five different algorithms [Bayesian algorithm (Munch *et al.* 2008a), BootstrapNJ algorithm (Munch *et al.* 2008a), BCM algorithm ('best close match', Meier *et al.* 2006), MD (minimum distance), MD + fuzzy membership (developed in this study)].

Appendix S3 (a–d) Detailed species assignments for Pacific Canadian marine fish (Steinke *et al.* 2009) based on COI sequences for 500 replications of queries (100 for Bayesian and BootstrapNJ methods) with five different algorithms [Bayesian algorithm (Munch *et al.* 2008a), BootstrapNJ algorithm (Munch *et al.* 2008a), BCM algorithm ('best close match', Meier *et al.* 2006), MD (minimum distance), MD + fuzzy membership (developed in this study)].

Appendix S4 (a–d) Detailed species assignments for temperate Europe butterfly (Dinca *et al.* 2011) based on COI sequences for 500 replications of queries (100 for Bayesian and BootstrapNJ methods) with five different algorithms [Bayesian algorithm (Munch *et al.* 2008a), BootstrapNJ algorithm (Munch *et al.* 2008a), BCM algorithm ('best close match', Meier *et al.* 2006), MD (minimum distance), MD + fuzzy membership (developed in this study)].

Appendix S5 (a–d) Detailed species assignments for 449 species of Diptera (Meier *et al.* 2006) based on COI sequences for 500 replications of queries (100 for Bayesian and BootstrapNJ methods) with five different algorithms [Bayesian algorithm (Munch *et al.* 2008a), BootstrapNJ algorithm (Munch *et al.* 2008a), BCM algorithm ('best close match', Meier *et al.* 2006), MD (minimum distance), MD + fuzzy membership (developed in this study)].

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.