

DNA Barcode-Based Molecular Identification System for Fish Species

Sungmin Kim¹, Hae-Seok Eo², Hyeyoung Koo³, Jun-Kil Choi³, and Won Kim^{1,4,*}

In this study, we applied DNA barcoding to identify species using short DNA sequence analysis. We examined the utility of DNA barcoding by identifying 53 Korean freshwater fish species, 233 other freshwater fish species, and 1339 saltwater fish species. We successfully developed a web-based molecular identification system for fish (MISF) using a profile hidden Markov model. MISF facilitates efficient and reliable species identification, overcoming the limitations of conventional taxonomic approaches. MISF is freely accessible at <http://bioinfosys.snu.ac.kr:8080/MISF/misf.jsp>.

INTRODUCTION

DNA barcoding is a molecular diagnostic method that provides rapid and accurate species identification (Hebert and Gregory, 2005; Schindel and Miller, 2005). Since the late 1990s, DNA sequences have been used to infer evolutionary relationships among various biological species (Woese and Fox, 1997), and DNA-based taxonomic approaches are now arguably a necessity in biodiversity studies (Tautz et al., 2002). Analysis of variations in short DNA sequences can accurately identify biological species (Hebert et al., 2003). Several studies have confirmed that a 648-bp segment of the 5' region of mitochondrial cytochrome *c* oxidase subunit I (*COI*) can be used as a DNA barcode to identify most animal species (Hebert et al., 2003; Ward et al., 2009). DNA barcoding facilitates rapid (Hajibabaei et al., 2006) and credible identification of immature specimens (Stahls et al., 2009) and closely related species (Rock et al., 2008). Further, a combination of morphological and molecular identification methodologies provides a practical approach for the discovery of new species (Hebert and Gregory, 2005).

While the development of automated DNA barcode identification libraries has been important for ecosystem research and conservation, several problems arise during the construction of a successful DNA barcoding system. First, incorrect species identification can be caused by unbalanced taxonomic sampling from within and among populations (Lefebure et al., 2006), despite the fact that there are large assemblages of DNA barcode sequences in the Barcode of Life Data Systems (BOLD,

<http://www.boldsystems.org>) (Ratnasingham and Hebert, 2007) and the Korea Barcode of Life database (KBOL, <http://korea-barcode.org>). To combat this problem, stringent statistical analyses should be performed to confirm species assignment (Nielsen and Matz, 2006). Second, complex evolutionary distance models or likelihood methods can generate obscure species groupings, and these require considerable computation time (Nielsen and Matz, 2006). Finally, unclear statistical significance values can be problematic when using the BOLD, TaxI, and BLAST systems (Ekrem et al., 2007; Steinke et al., 2005).

In the present study, we adopted a machine learning approach to resolve these problems (Baldi and Brunak, 2001). We applied this approach to biological datasets consisting of the sequences from 53 Korean freshwater fish species, 233 other freshwater fish species, and 1339 saltwater fish species. We developed a web-based Molecular Identification System for Fish (MISF) based on a profile hidden Markov model (HMM) that is suitable for modeling species-specific sequence patterns (Eddy, 1998). We conclude that the MISF can be used to correctly identify fish species. Moreover, this is the first study in which a statistical HMM algorithm approach has been successfully used to identify fish species.

MATERIALS AND METHODS

DNA sequence dataset

We collected 398 specimens from streams and rivers in South Korea. Genomic DNA was extracted from all fish using the Qiagen DNeasy Blood and Tissue kit, and the *COI* region was sequenced using PCR with the Folmer primer set (Folmer et al., 1994). All PCR reactions were performed using the following thermal cycling program: 1 min at 94°C followed by 35 cycles of 0.5 min at 94°C, 1 min at 50°C, and 1.5 min at 72°C, and a final extension for 10 min at 72°C. The 25 μ l-PCR reaction mix consisted of 14.7 μ l ultrapure water, 5 μ l 5 \times PCR buffer, 1 μ l of each primer (10 μ M), 1 μ l dNTP (10 mM), 0.3 μ l Taq polymerase (5U), and 2 μ l genomic DNA template. All sequences have been deposited in the Korea Barcode of Life (KBOL) database, along with specimen details, including date of collection, specimen source, and specimen images. These data can be found in the "Korean freshwater fish 2007-2009" project, which can

¹Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea, ²Information & Technology Laboratory, LG Electronics Inc., Seoul 137-130, Korea, ³Department of Biological Science, Sangji University, Wonju 220-702, Korea, ⁴School of Biological Sciences, Seoul National University, Seoul 151-742, Korea

*Correspondence: wonkim@plaza.snu.ac.kr

be accessed via the Model ecosystems (KMOD) link on the KBOL website.

In addition, 7,972 *COI* sequences from 1,572 fish species were used for our analyses. These sequences were downloaded from the 24 representative projects under the FishBOL category in the Barcode of Life Data System (BOLD). Sequence divergence was quantified using the Kimura 2-parameter (K2P) distance model (Kimura, 1980), and neighbor-joining (NJ) trees were generated using MEGA version 4.0.2 (Tamura et al., 2007).

Building profile hidden Markov models

Sequence variation within a species varies from species to species. Thus, we sorted sequences of each species according to sequence similarity scores as calculated by MUSCLE v3.8 software, a multiple alignment program (Edgar, 2004). If the number of different sequences within a species was less than four, we repeatedly generated the same sequences. Statistical models were constructed using the *hmmbuild* and *hmmcalibrate* programs from HMMER v2.3.2 software (Eddy, 1998; 2008). Emission and transition probabilities were calculated by averaging four candidate sequences for each species (Supplementary Table 3). The *hmmpfam* program applied to the above described libraries was used to identify species-specific patterns of unknown DNA barcode sequences. The *blastall* program was used to execute BLAST (Altschul et al., 1990) in order to achieve rapid sequence alignment of unknown query sequences and sequences from reference barcode databases. The results are listed in ascending order according to the match score and E-value (expectation value).

Optimum conditions and cutoff value

All steps were automatically processed by Perl scripts and controlled by cutoff values. Cutoff values indicate the standard deviation of HMMER bit scores when constructing a profile HMM. The HMMER bit scores indicate whether the sequence is a better match to the profile model or to the null model of non-homologous sequences as follows (Eddy, 2008):

$$S = \log_2 \frac{P(seq | HMM)}{P(seq | null)}, \quad \sigma = \sqrt{\frac{\sum (S - \bar{S})^2}{N}}, \quad (1)$$

where N is the number of candidate sequences trained for building a profile HMM for each species, S is the HMMER bit score, and σ is the standard deviation value of the HMMER bit scores.

The cutoff value filters out candidate sequences that are used for multiple sequence alignments during the model-building process. Candidate sequences were repeatedly compared with other sequences in the same range (0.5-3.0, with an interval of 0.1) to determine the best conditions with minimum type I errors. We consequently found the best cutoff value (1.0, Fig. 2) and determined the best conditions (C) when building a model as follows:

$$C = \begin{cases} N \leq 3, \text{ Same sequences are randomly generated till} \\ \quad N = 4 \\ N \geq 4, N \text{ is increased under cutoff value condition,} \\ \quad 0 \leq \sigma \leq 1.0, \end{cases} \quad (2)$$

where N is the number of sequences and σ is the standard deviation of the HMMER bit scores.

Validation and performance evaluation

We adopted K-fold and leave-one-out cross validation (LOOCV) approaches to evaluate the performance of profile HMMs. LOOCV requires more computation than K-fold cross validation; however, LOOCV identifies all errors caused by imbalanced training datasets (Meyer and Paulay, 2005). Under the LOOCV approach, a profile HMM is programmed by leaving out each individual sequence of the training set in each round. A true error rate, E , is computed as the average of the individual error rates as follows:

$$E = \frac{1}{n} \sum_{i=1}^n E_i, \quad (3)$$

where n is the number of iterations and E_i is the individual error rates.

We used a variety of standard measurements to evaluate the quality of prediction performances, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy (Eo et al., 2009).

Web-based molecular identification system for fish (MISF)

MISF is a web-based identification system that uses both BLAST and profile HMM algorithms. We developed a web-based MISF that was constructed in Java-based JSP and operated on the Apache Tomcat web-server in Windows XP. The MISF is currently optimized for IE7 and ChromePlus v1.4.1.0 internet browsers. This molecular diagnostic method is freely available to all users at <http://bioinfosys.snu.ac.kr:8080/MISF/misf.jsp> (Supplementary Fig. 3).

RESULTS

Sequence analysis of 1,625 fish species was implemented using large datasets, including 53 Korean freshwater fish (KFW), 233 other freshwater fish (FW), and 1339 saltwater fish (SW) (Supplementary Table 1).

Sequence recoverability and NUMTs in the KFW dataset

Three hundred and ninety-eight novel *COI* barcodes were generated from the 53 species. These barcodes represented 24.9% of the 213 Korean freshwater fish species. When PCR conditions were optimized, all samples were easily amplified. The mitochondrial *COI* region had a high overall rate of recovery (93.87%, Table 1). During initial barcode amplification attempts, nuclear copies of mitochondrial DNA (NUMTs) were occasionally co-amplified. Nuclear co-amplification can interfere with *COI* barcode analysis (France and Hoover, 2002). Thus, it was necessary to filter the *COI* barcodes by careful examination of sequence characteristics, such as indels and in-frame stop codons (Bensasson et al., 2001; Song et al., 2008). The protocol outlined in Supplementary Fig. 1 describes how to avoid and identify NUMTs. As shown in Table 1, 21 NUMTs, including in-frame stop codons ($n = 19$) and deletions ($n = 2$), were discovered and excluded from the dataset. We also performed credible species identification for 28 immature specimens and were able to correct 22 cases of species misidentification caused by morphologically ambiguous traits.

Intraspecific and interspecific variation in fish species

The sequence analysis data from the three datasets were used to assess the efficiency of *COI* barcodes in species identification (Supplementary Table 2). Both intraspecific and interspecific distances were assessed using K2P. Maximum variations within species *COI* sequences were on average 0.48%, 0.38%,

Table 1. Summary of sample information

	Successfully sequenced samples	NUMTs*	Low-quality sequences	Samples reidentified by DNA barcoding	Immature specimens (larvae and eggs)
No. of samples/total	398/424	21/424	5/424	22/398	28/398
percentage (%)	93.87	4.95	1.18	5.53	8.42

NUMTs*: *COI* nuclear mitochondrial pseudogenes

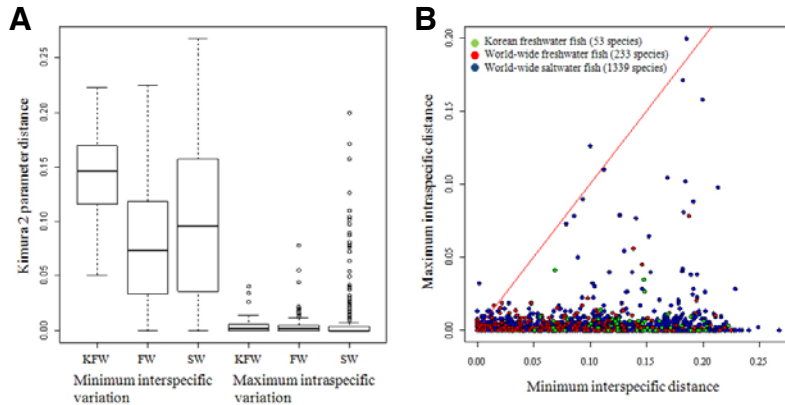


Fig. 1. (A) Box plots depicting the substantial difference between minimum interspecific variation and maximum intraspecific variation using K2P distance modeling of sequence divergence. (B) DNA barcode gap. Minimum interspecific and maximum intraspecific sequence divergence values were calculated to evaluate *COI* as a DNA barcode for species determination. Most species fall below the 1:1 line (red line), indicating that *COI* is a highly suitable barcode.

and 0.37%, whereas minimum variations among species averaged 13.96%, 7.89%, and 9.75% for the datasets KFW, FW, and SW, respectively. Under a stricter statistical criterion proposed by Hebert (Hebert et al., 2004), we determined that the differences between maximum intraspecific and minimum interspecific variation in the three datasets were 28.78, 20.59, and 26.07 times, respectively (Fig. 1A). These differences indicate that there is considerable genetic variation among species. There were, however, some notable erroneous cases: minimum interspecific variation was zero in 69 (4.25%, $n = 6$ in FW and $n = 63$ in SW) of the 1,625 species and overlap was observed in 96 (5.90%, $n = 11$ in FW and $n = 85$ in SW) of the 1,625 species. However, most *COI* barcodes were observed below the 1:1 line representing a species with a barcode gap, suggesting that with the exception of 28 cases (6 species in FW and 22 species in SW), the barcodes are highly suitable for species identification (Fig. 1B).

Optimal cutoff value and HMMs

Cutoff values are critical when constructing a statistical library since the proposed cutoff value provides an alternative criterion for screening candidate sequences. We assessed the false-positive error rates of the test datasets and concluded that the optimal cutoff value is 1.0 (Fig. 2). The profile HMMs based on the KFW were validated under LOOCV testing conditions, whereas two other profile HMMs, based on FW and SW, were evaluated with three-fold cross validation because of their large sample size. The profile HMMs constructed from KFW achieved the best accuracy, and a zero standard error rate was calculated at both genus and species levels (Table 2). This result was corroborated by analysis of the negative dataset from the Barcoding of Canadian Freshwater fishes project (BCF; Hubert et al., 2008, BOLD system). The remaining two models also displayed good prediction accuracy at the genus level (> 97%) and at the species level (> 93%) (Table 3). False positives can result from the absence of a barcode gap and unbalanced sampling as opposed to methodological limitations, as illustrated by the 28 exceptional cases shown in Fig. 1B. Although the number of sequences within species varies (Supplementary

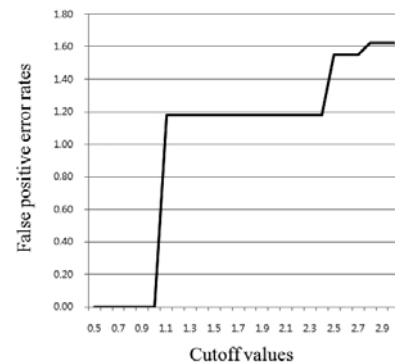


Fig. 2. To select an optimal cutoff value, false-positive error rates were calculated in each step of the range between 0.5 and 3.0. False positive error rates remained at zero until a cutoff value of 1.0, were sustained at 1.18 for cutoff values of 1.1 to 2.3, and then gradually increased at a cutoff value of 2.4 and above.

Table 3), sampling using the best cutoff value and a strategy using the HMM showed that the statistical library can provide clear species differentiation without over- or underestimation. A verification procedure for barcode sequences and the application of these sequences is illustrated in Supplementary Fig. 1.

Character-based discrimination of closely related species

Species identity was wrongly assigned to some individuals using HMM searches (Table 3) and the NJ tree method (Supplementary Fig. 2). Some were categorized as closely related species with no barcode gap (Fig. 1B). In addition, a coalescent-based statistical method was applied to differentiate the species with a considerable number of fixed nucleotide substitutions (Tavares and Baker, 2008). The results were obtained using the *randomForest* package of R programming language (Breiman, 2001; Ihaka and Gentleman, 1996). This type of analysis showed that unique patterns of nucleotide composition exist among species. Among the monophyletic clusters, 21

Table 2. Performance comparisons for models defined in a range of threshold values for each HMM. Each model was evaluated using the same test set consisting of 398 positive and 1338 negative freshwater fish sequences.

HMM cutoff value	True positive	False positive	True negative	False negative	Sensitivity	Specificity	PPV ¹	NPV ²	Accuracy (%)
$0 \leq \sigma \leq 1.0$	398	0	1338	0	1	1.0000	1.0000	1	100.0000
$1.1 \leq \sigma \leq 2.4$	382	16	1338	0	1	0.9882	0.9598	1	99.0783
$2.5 \leq \sigma \leq 2.7$	377	21	1338	0	1	0.9845	0.9472	1	98.7903
$2.8 \leq \sigma \leq 3.0$	376	22	1338	0	1	0.9838	0.9447	1	98.7327

¹Positive predictive value²Negative predictive value**Table 3.** Typical datasets from HMMs and performance levels

Dataset	No. of individuals	No. of species	No. of genera	Accuracy at genus level (%)	Accuracy at species level (%)
KFW ¹	398	53	41	100	100
FW ²	1575	233	111	99.11	95.68
SW ³	6397	1339	702	97.56	93.79

¹KFW, Korean freshwater fish²FW, World-wide freshwater fish³SW, World-wide saltwater fish**Table 4.** Estimated results using the R package *randomForest*

Dataset	Genus	No. of individuals	No. of species	No. of parsimony-informative sites	No. of variables tried at each split	OOB* estimate of error rate (%)
Monophyletic groups	<i>Bramocharax</i>	51	2	19	7	0
	<i>Catostomus</i>	38	4	85	9	0
	<i>Ichthyomyzon</i>	9	2	33	5	0
	<i>Oncorhynchus</i>	49	7	119	10	0
	<i>Rhinichthys</i>	43	6	84	9	0
	<i>Amblyraja</i>	39	4	26	5	2.56
	<i>Notropis</i>	104	14	182	13	2.73
	<i>Atlantoraja</i>	58	3	59	7	3.45
	<i>Platycephalus</i>	40	9	201	14	5
	<i>Raja</i>	255	11	167	12	10.98
	<i>Carcharhinus</i>	269	23	141	11	11.15
	<i>Rajella</i>	23	6	167	9	17.39
	<i>Hippocampus</i>	56	23	173	13	25
Paraphyletic groups	23 genus groups including 6 species in FW and 22 species in SW (no barcode gap)	1239	147	273	16	12.19
KFW	41 genus groups	398	53	280	16	1.51
FW	111 genus groups	1575	233	285	16	3.68
SW	702 genus groups	6397	1339	315	17	11.7

OOB* (Out-of-bag data): This is used to obtain a running unbiased estimation of classification error as trees are added to the forest.

species of five genera (*Bramocharax*, *Catostomus*, *Ichthyomyzon*, *Oncorhynchus*, and *Rhinichthys*) were perfectly classified (Table 4). Low error rates were observed in *Amblyraja*, *Notropis*, *Atlantoraja*, and *Platycephalus*, although these genera have many parsimony-informative sites. Unacceptably high DNA barcode error rates were observed in the remaining four genera

(*Raja*, *Carcharhinus*, *Rajella*, and *Hippocampus*). Twenty-three genera, including 28 species that have no barcode gap, also generated considerable error rates (Table 4). We also analyzed three large datasets using the same methods and obtained better results, although numerous species groups and overlapping parsimony-informative sites were still observed (Table 4).

DISCUSSION

Use of HMM for breaking unbalanced sampling and measurement errors

MISF is the first species identification approach that utilizes profile HMMs. The profile HMMs were developed using the same criteria to avoid errors caused by unbalanced taxonomic sampling from within and among populations (Lefebure et al., 2006). The search performance showed that a significant number of false positives can be reduced by employing an optimal cutoff value (Table 2). MISF was highly accurate at both species and genus levels (Table 3). We showed that credible species models can be developed using HMM compared with other methods that rely on sequence similarity, such as BLAST, or distance models, such as Taxl software (Nielsen and Matz, 2006; Steinke et al., 2005). MISF provides clear statistical significance rather than potential erroneous probabilities generated by BOLD (Ekrem et al., 2007) and BLAST (Steinke et al., 2005). It is clear via cross validation that MISF allows for high statistical significance (Baldi and Brunak, 2001). By sequence analysis, we found that the ratio of no overlap between minimum interspecific and maximum intraspecific distance to the total number of sequences was 95.27% and 93.65% for FW and SW, respectively. However, we obtained a higher accuracy at the species level (Table 3). All in all, MISF is fast, easy to use, and will be very useful in the fields of systematics, evolutionary biology, and ecology.

Diagnostic approach for closely related species

Currently, several approaches are used to discriminate between closely related species. Character-based approaches in particular have suggested that the *COI* barcode is useful for species determination (Tavares and Baker, 2008). Diagnostic characters resulting in species clusters have some advantages. Fixed nucleotide substitutions from *COI* sequences can be extracted without recourse to evolution models or consideration of a significant cutoff value (Vaugh, 2007). By employing a strict threshold (Fig. 1B), we discovered that there was no barcode gap or zero minimum interspecific variation between closely related species (Supplementary Table 2). In a previous study, individual character states were not evaluated using statistical analyses (Rach et al., 2008; Tavares and Baker, 2008). In the present study, we evaluated priority rankings for character placement in tree branches (Supplementary Fig. 4). This character-based diagnostic approach can be used as an alternative approach for differentiating closely related species; however, there is a degree of error associated with this method (Table 4). DNA barcoding based on only a single gene (*COI*) is somewhat controversial in the taxonomy community. Some erroneous results have emphasized the limitations of single gene *COI* barcoding as opposed to bioinformatics and biological approaches (see Results, Baker et al., 2009).

Future direction of MISF

Our study shows that MISF is an attractive system for identifying fish species. This approach will be extensively applied to the KBOL database system, which comprises a collection of biodiversity data along with DNA barcodes. The molecular identification system will definitely leverage researchers to directly link DNA barcodes and biodiversity information and to undertake conservation activities, including monitoring and detection of specific organisms. Future species models are likely to be more elaborate and powerful because mtDNA barcoding will be supplemented with nuclear barcoding to facilitate greater reliability (Austerlitz et al., 2009). Moreover, taxonomic approaches that

integrate DNA sequencing, morphology, and ecological studies will achieve maximum species identification efficiency. The automated identification system MISF will certainly facilitate fast and reliable species identification to promote the integration of a variety of biological information.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This work was supported by a grant from the Eco-Technopia 21 Project, which is funded by the Ministry of Environment of Korea (2008-05002-0065-0); the second stage of the Brain Korea 21 Project in 2009; and a 2008-2009 Sangji University Sabbatical Leave Grant.

REFERENCES

- Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., and Laredo, C. (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics* 10, S10.
- Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.
- Baker, A., Tavares, E., and Elbourne, R. (2009). Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Mol. Ecol. Res.* 9, 257-268.
- Baldi, P., and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. The MIT Press.
- Bensasson, D., Zhang, D., Hartl, D., and Hewitt, G. (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* 16, 314-321.
- Breiman, L. (2001). Random forests. *Machine Learning* 45, 5-32.
- Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755-763.
- Eddy, S. (2008). HMMER user's guide. Department of Genetics, Washington University School of Medicine 2.
- Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792-1797.
- Ekrem, T., Willassen, E., and Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Mol. Phylogenet. Evol.* 43, 530-542.
- EO, H., Kim, S., Koo, H., and Kim, W. (2009). A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins. *Mol. Cells* 27, 629-634.
- Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Marine Biol. Biotech.* 3, 294-299.
- France, S., and Hoover, L. (2002). DNA sequences of the mitochondrial *COI* gene have low levels of divergence among deep-sea octocorals (Cnidaria: Anthozoa). *Hydrobiologia* 471, 149-155.
- Hajibabaei, M., Janzen, D., Burns, J., Hallwachs, W., and Hebert, P. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proc. Natl. Acad. Sci. USA* 103, 968-971.
- Hebert, P., and Gregory, T. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biol.* 54, 852-859.
- Hebert, P., Ratnasingham, S., and deWaard, J. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Royal Soc. London-B* 270, 96-99.
- Hebert, P., Stoeckle, M., Zemlak, T., and Francis, C. (2004). Identification of birds through DNA barcodes. *PLoS Biol.* 2, 1657-1663.
- Hubert, N., Hanner, R., Holm, E., Mandrak, N., Taylor, E., Burrige, M., Watkinson, D., Dumont, P., Curry, A., and Bentzen, P. (2008). Identifying Canadian freshwater fishes through DNA barcodes. *PLoS ONE* 3, e2490.
- Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statistics* 5, 299-314.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111-120.
- Lefebure, T., Douady, C., Gouy, M., and Gibert, J. (2006). Relation-

- ship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. *Mol. Phylogenet. Evol.* *40*, 435-447.
- Meyer, C., and Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* *3*, e422.
- Nielsen, R., and Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biol.* *55*, 162-169.
- Rach, J., DeSalle, R., Sarkar, I., Schierwater, B., and Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc. Royal Soc. B* *275*, 237-247.
- Ratnasingham, S., and Hebert, P. (2007). BOLD: the Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* *7*, 355-364.
- Rock, J., Costa, F., Walker, D., North, A., Hutchinson, W., and Carvalho, G. (2008). DNA barcodes of fish of the Scotia Sea, Antarctica indicate priority groups for taxonomic and systematics focus. *Antarctic Sci.* *20*, 253-262.
- Schindel, D., and Miller, S. (2005). DNA barcoding a useful tool for taxonomists. *Nature* *17*, 435.
- Song, H., Buhay, J., Whiting, M., and Crandall, K. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. USA* *105*, 13486-13491.
- Stahls, G., Vujic, A., Perez-banon, C., Radenkovic, S., Rojo, S., and Petanidou, T. (2009). *COI* barcodes for identification of Merodon hoverflies (Diptera, Syrphidae) of Lesvos Island, Greece. *Mol. Ecol. Res.* *9*, 1431-1438.
- Steinke, D., Vences, M., Salzburger, W., and Meyer, A. (2005). Taxl: a software tool for DNA barcoding using distance methods. *Philos. Transact. B* *360*, 1975-1980.
- Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* *24*, 1596-1599.
- Tautz, D., Arctander, P., Minelli, A., Thomas, R., and Vogler, A. (2002). DNA points the way ahead in taxonomy. *Nature* *418*, 479-479.
- Tavares, E., and Baker, A. (2008) Single mitochondrial gene barcodes reliably identify sister-species in diverse clades of birds. *BMC Evol. Biol.* *8*, 81.
- Ward, R., Hanner, R., and Hebert, P. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *J. Fish Biol.* *74*, 329-356.
- Waugh, J. (2007). DNA barcoding in animal species: progress, potential and pitfalls. *BioEssays* *29*, 188-197.
- Woese, C., and Fox, G. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* *74*, 5088-5090.