THE ROYAL SOCIETY biology letters

Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species

Paul D. N. Hebert*, Sujeevan Ratnasingham and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada
** Author for correspondence (phebert@uoguelph.ca).*

Recd 09.03.03; *Accptd* 28.03.03; *Online* 15.05.03

With millions of species and their life-stage transformations, the animal kingdom provides a challenging target for taxonomy. Recent work has suggested that a DNA-based identification system, founded on the mitochondrial gene, cytochrome *c* oxidase subunit 1 (COI), can aid the resolution of this diversity. While past work has validated the ability of COI sequences to diagnose species in certain taxonomic groups, the present study extends these analyses across the animal kingdom. The results indicate that sequence divergences at COI regularly enable the discrimination of closely allied species in all animal phyla except the Cnidaria. This success in species diagnosis reflects both the high rates of sequence change at COI in most animal groups and constraints on intraspecific mitochondrial DNA divergence arising, at least in part, through selective sweeps mediated via interactions with the nuclear genome.

Keywords: molecular taxonomy; DNA barcode; cytochrome *c* oxidase subunit 1; DNA; mitochondrial

1. INTRODUCTION

A final taxonomic system for the animal kingdom will probably include at least 10 million species partitioned among more than a million genera. Given such high diversity, there is a growing realization that it is critical to seek technological assistance for its initial description and its subsequent recognition (Godfray 2002; Blaxter 2003). Recent investigations have suggested the feasibility of creating identification systems reliant on the analysis of sequence diversity in small segments of DNA (Tautz *et al.* 2003). Hebert *et al.* (2003) focused this discussion by proposing that a DNA barcoding system for animal life could be based upon sequence diversity in cytochrome *c* oxidase subunit 1 (COI). They established that diversity in the amino acid sequences coded by the 5′ section of this mitochondrial gene was sufficient to reliably place species into higher taxonomic categories (from phyla to orders). They also found that diversity in nucleotide sequences of the same gene region regularly permitted the discrimination of closely allied species of lepidopterans, a group with modest rates of molecular evolution and high species diversity. As such, these insects provided a challenging test for the ability of COI diversity to resolve species boundaries.

Although Hebert *et al.* (2003) argued that a COI-based identification system could be developed for all animals, scepticism has been expressed (Mallet & Willmot 2003). Primary objections have focused on the concern that DNA sequence differences among closely allied species will often be too small to allow their discrimination. Although this issue has never been tested comprehensively, Johns & Avise (1998) demonstrated that closely related species of vertebrates regularly show more than 2% divergence at another mitochondrial gene, cytochrome *b*. The present study addresses this issue further by examining the extent of sequence diversity at COI among congeneric taxa in the major animal phyla. The most intensive analysis focuses on the arthropods because sequence information for these organisms is particularly detailed owing to their high taxonomic diversity. However, COI divergences are also examined among closely related species in all animal phyla where data are available. In total, sequence divergences are examined in more than 13 000 congeneric pairs including representatives from 11 phyla. These results support, with the exception of a single phylum, the conclusion that species-level diagnoses can routinely be obtained through COI analysis.

2. MATERIAL AND METHODS

Sequences were extracted from GenBank between November 2002 and February 2003 for all congeneric species pairs of animals possessing at least 400 bp of COI sequence from homologous sites. In total, 2238 species met this criterion. Because their COI sequences were acquired using varied primers, they derive from different sections of the gene. In practice, most comparisons involved either a sequence block that extended from near the 5′ end of the gene to its middle or a block extending from the midst of the gene to its 3′ end. Rates of sequence change in the 5′ and 3′ halves of COI were compared by examining divergences in these regions for all pairwise comparisons of the 260 animal species with full mitochondrial genomes in GenBank release 131.0. This analysis (results not shown) indicated that sequence divergences in the halves were closely similar (mean sequence divergence in COI-5′ was 97.7% of that in the 3′ region, s.d. 6.2%). Because of this congruence, the measures of sequence divergence for other species pairs are analysed without reference to their source region in the gene.

Following the extraction of COI sequences from GenBank, *p*-distances ($p = n_d/n_t$ where n_d is the number of different nucleotides between the two sequences and n_t is the total number of nucleotides examined) were determined for each congeneric species pair. A single divergence value was obtained for all genera represented by two species, while $n(n-1)/2$ values were obtained for genera with three or more species. The large number of sequences available for congeneric species pairs in the Hexapoda permitted separate analysis for the four dominant orders (Coleoptera, Diptera, Lepidoptera and Hymenoptera), while the remaining taxa were pooled as 'other hexapods'. Data for the chelicerates and crustaceans were more limited, so divergence values for these arthropods were not partitioned to the ordinal level. Because less information was available for the Annelida, Chordata, Cnidaria, Echinodermata, Mollusca, Nematoda and Platyhelminthes, these data were aggregated at a phylum level. Finally, results for three small phyla (Bryozoa, Brachiopoda and Onychophora) were pooled. An alignment for all of the animal COI sequences in GenBank is available at www.uoguelph.ca/~phebert. As well, a list of the 447 genera examined in the present analysis, together with sequence divergences (mean, maximum, minimum) for the species pairs in each genus, is included in electronic Appendix A, available on The Royal Society's Publications Web site.

3. RESULTS

COI divergences among the 13 320 species pairs ranged from a low of 0.0% to a high of 53.7% (figure 1). While most pairs (79%) showed greater than 8% sequence divergence, levels did vary among higher taxonomic groups

DOI 10.1098/rsbl.2003.0025

Table 1. Mean and standard deviation of the percentage sequence divergences at COI for 13 320 congeneric species pairs in 11 animal phyla.
(The percentage of sequence divergence estimates falling in a particular range is also shown. *n* indicates the number of congeneric pairs examined in each group.)

				COI sequence divergence (%)						
phylum	*n*	mean	s.d.	0–1	1–2	2–4	4–8	8–16	16–32	32+
Annelida	128	15.7	6.2	6.3	1.6	—	3.9	18.0	70.3	—
Arthropoda										
Chelicerata	1249	14.4	3.6	—	0.2	0.2	2.0	50.8	46.8	—
Crustacea	1781	15.4	6.6	0.1	0.3	4.3	13.4	18.0	63.8	0.1
Coleoptera	891	11.2	3.8	2.2	1.6	3.0	8.0	74.2	11.0	—
Diptera	1429	9.3	3.5	0.9	2.1	4.1	14.0	76.2	2.7	—
Hymenoptera	2993	11.5	3.8	0.2	—	0.3	3.3	93.0	3.2	—
Lepidoptera	882	6.6	2.2	1.0	2.8	7.3	60.4	28.5	—	—
other orders	1458	10.1	4.9	0.5	1.6	8.4	35.5	41.8	12.1	—
Chordata	964	9.6	3.8	1.2	0.7	4.3	19.2	61.7	12.9	—
Cnidaria	17	1.0	1.2	88.2	5.9	5.9	—	—	—	—
Echinodermata	86	10.9	4.9	1.2	1.2	5.8	39.5	44.2	8.1	—
Mollusca	1155	11.1	5.1	1.2	1.9	4.0	15.0	67.5	10.0	0.4
Nematoda	49	11.0	2.9	—	2.0	—	22.4	73.5	2.0	—
Platyhelminthes	84	14.4	5.4	8.3	—	—	4.8	44.0	42.9	—
minor phyla	154	13.3	9.7	0.6	1.3	2.6	39.6	38.3	16.9	0.7
overall	13 320	11.3	5.3	0.9	1.0	3.4	16.2	59.4	19.0	0.1

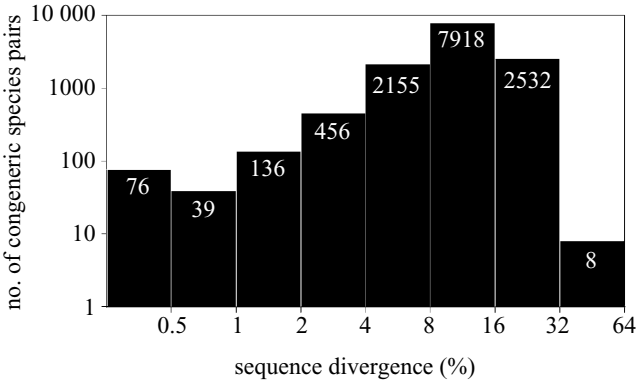

Figure 1. COI sequence divergences for 13 320 congeneric pairs of animal species belonging to 11 phyla.

(table 1). The Cnidaria showed far lower COI divergences than any other phylum with 94.1% of the cnidarian pairs possessing less than 2% sequence divergence, while just 1.9% of the species pairs in other taxonomic groups showed such limited differentiation. This difference was highly significant ($G = 489.55$, $p < 0.0001$).

Levels of divergence varied among the four major insect orders with lepidopterans showing a lower average sequence divergence than the other three (table 1). Divergences in the insects, as a group, were less than those in the other two dominant groups of arthropods, the chelicerates and crustaceans. Mean divergences in the other major phyla were comparable to the overall mean divergence value for the arthropods. Sample sizes were small for the minor phyla, but their mean level of divergence appeared similar to those in the large phyla.

4. DISCUSSION

This study has established that congeneric species of animals regularly possess substantial sequence divergence in their COI genes. In fact, more than 98% of species pairs showed greater than 2% sequence divergence. The few cnidarians included in the analysis showed far less divergence than that typical for other animal phyla. This result is not surprising as rates of mitochondrial evolution are exceptionally low in these organisms (France & Hoover 2002; Shearer *et al.* 2002). Their stasis seems linked, at least in part, to the presence of an excision repair system absent in other animal mitochondria. Low rates of mitochondrial divergence also appear typical of the plant kingdom (Palmer 1992). By contrast, levels of COI divergence in the Fungi and Protista exceed those in animals (Ingaki *et al.* 1998), suggesting that the prospects for extending a COI-based identification system to these kingdoms are high.

Except for the cnidarians, most congeneric species pairs of animals showed enough sequence divergence to ensure their easy diagnosis. In fact, the mean divergence value of 11.3% indicates that most pairs were separated by more than 50 diagnostic substitutions in every 500 bp of their COI gene. About 1.9% of the congeneric pairs in groups with normal rates of mitochondrial evolution showed less than 2% divergence, probably often reflecting short histories of reproductive isolation. Such pairs may be separable, even in samples from natural populations where intraspecific variation might be expected to cloud decisions. For example, lepidopteran species pairs with similarly low levels of COI divergence were invariably resolved (Hebert *et al.* 2003). Some additional cases of low divergence may simply be artefacts generated by flawed identifications, but other cases of congruence will undoubtedly reflect mitochondrial introgression. In the latter situations, the analysis of a rapidly evolving nuclear sequence, such as the internal transcribed spacer region of the ribosomal repeat, will aid taxonomic resolution.

Although more work needs to be done to quantify levels of COI divergence within animal species, a good sense of

the usual limits of intraspecific divergence in mitochondrial genes derives from phylogeographic analyses. Two results of these studies are particularly important. First, intraspecific divergences are rarely greater than 2% and most are less than 1% (Avise 2000). Second, when higher divergences are observed, these variants ordinarily occur as geographical isolates, reflecting their origin in past episodes of gene pool fragmentation. Moreover, many of these high divergences involve cases of taxonomic uncertainty where lineages share a species epithet, but their actual status is unclear (Avise & Walker 1999). Collectively, these phylogeographic studies do establish that intraspecific divergences are ordinarily well below those that separate congeneric species pairs. Hence, COI divergences can serve as an effective tool in species recognition. Moreover, the fact that some 'species' show higher divergences does not compromise the use of COI sequences for their identification. In fact, just the opposite, it allows delineation of the regional lineages that comprise them.

Concern has been expressed that efforts to base identification systems on mtDNA markers would fail because of the prevalence of horizontal transfers of mitochondria between divergent lineages and because closely allied species would regularly share mitochondrial polymorphisms that were millions of years old (Mallet & Willmot 2003). The present study suggests that such complications are rare. In fact, the clear delineation of most congeneric species pairs indicates a surprising ferocity of lineage pruning. The restricted levels of intraspecific mitochondrial divergence that result from this pruning are critical to taxon diagnosis and they may have a simple explanation. Horizontal transfers of mtDNA and the persistence of ancestral polymorphisms require the autonomy of mitochondrial genomes. But, there is growing evidence that mitochondrial and nuclear genomes are linked in a *pas de deux* of surprising intimacy, best evidenced by studies on cybrids. For example, despite the close genetic similarity of their nuclear genomes, orang-utan mitochondria experience a total collapse in respiratory capacity when placed in a human cell background (Barrientos *et al.* 2000). Even chimpanzee and gorilla mitochondria show 20% reductions in their oxidative capacity in a human cytogenetic setting (Barrientos *et al.* 1998). These effects have been tied to interactions between gene products encoded by the nucleus and the mitochondrion; the three mitochondrial enzymes in the cytochrome *c* oxidase system are decorated by 10 nuclear proteins critical to their functionality (Wu *et al.* 2000). Similar nuclear–cytoplasmic incompatibilities have been detected in studies that employed recurrent backcrosses to transfer mitochondria into a 'foreign' nuclear background. For example, backcrosses between regional groups of the copepod *Tigriopus californicus* lead to offspring whose respiratory capacity is compromised (Willett & Burton 2001). Detailed molecular studies have revealed that selection acts particularly strongly to reconfigure mitochondrial gene products, apparently to optimize their interactions with nuclear 'accessory' proteins whose amino acid composition is far more static. Moreover, the mitochondrial substitutions are tightly targeted. All three cytochrome *c* oxidase genes in primate mitochondria show elevated levels of non-synonymous substitutions at amino acid positions that are closely juxtaposed to nuclear-encoded proteins (Schmidt *et al.* 2001). Comparative studies of the patterns of genetic diversity in mitochondrial and nuclear genomes have additionally provided evidence for the depletion of mtDNA diversity via selective sweeps, perhaps mediated by the rise of mitochondrial variants with more effective nuclear interactions (Ballard 2000; Gerber *et al.* 2001). These sweeps do not need to occur with high frequency to ensure the clear delineation of sister species. Given a mean species longevity of 5 million years, the occurrence of sweeps at average intervals of 500 000 years would eliminate ancestral polymorphisms while allowing the accumulation of large amounts of shallow diversity, a pattern congruent with that seen in nature. Interactions between mitochondrial and nuclear genomes should also ensure that horizontal transfers are ordinarily fatal. Of course, cases of mitochondrial transfer have been detected between some closely allied animal species (e.g. Glemet *et al.* 1998), but these may reflect situations in which key nuclear genes were also introgressed, restoring functionality to the misplaced mitochondria.

The present results indicate that an identification system for animal life based on the COI gene will be highly effective. Although COI divergences appear too low to regularly enable species diagnosis within the cnidarians, generic-level identifications in these organisms remain a prospect. More importantly, the mitochondrial genomes of closely allied species in other phyla, those that comprise the bulk of animal diversity, regularly show sufficient sequence diversity to enable their discrimination. Even if COI analysis simply generated robust generic assignments, its application would winnow the 10 million animal species down to a generic assemblage averaging less than 10 species, delivering a resolution of 99.9999% of animal diversity in the process. As the present study has established that species-level identifications are ordinarily achieved, COI analysis actually provides a taxonomic system that is chasing the last digit of animal diversity.

Acknowledgements

This work was supported by grants from NSERC and the Canada Research Chairs programme to P.D.N.H. We thank Teri Crease, Jinzhong Fu, Bob Ward and Jonathan Witt for their thoughtful comments on the manuscript.

Avise, J. C. 2000 *Phylogeography. The history and formation of species*. Cambridge, MA: Harvard University Press.

Avise, J. C. & Walker, D. 1999 Species realities and numbers in sexual vertebrates: perspectives from an asexually transmitted genome. *Proc. Natl Acad. Sci. USA* **96**, 992–995.

Ballard, J. W. O. 2000 Comparative genomics of mitochondrial DNA in *Drosophila simulans*. *J. Mol. Evol.* **51**, 64–75.

Barrientos, A., Kenyon, L. & Moraes, C. T. 1998 Human xenomitochondrial cybrids. Cellular models of mitochondrial complex I deficiency. *J. Biol. Chem.* **273**, 14 210–14 217.

Barrientos, A., Muller, S., Dey, R., Weinberg, J. & Moraes, C. T. 2000 Cytochrome *c* oxidase assembly in primates is sensitive to small evolutionary variations in amino acid sequence. *Mol. Biol. Evol.* **17**, 1508–1519.

Blaxter, M. 2003 Counting angels with DNA. *Nature* **421**, 122–124.

France, S. C. & Hoover, I. L. 2002 DNA sequences of the mitochondrial COI gene have low levels of divergence among deep-sea octocorals (Cnidaria: Anthozoa). *Hydrobiol.* **471**, 149–155.

Gerber, A. S., Loggins, R., Kumar, S. & Dowling, T. E. 2001 Does non-neutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? *A. Rev. Genet.* **35**, 539–566.

Glemet, H., Blier, P. & Bernatchez, L. 1998 Geographical extent of arctic char (*Salvelinus alpinus*) mtDNA introgression in brook char populations (*Salvelinus fontinalis*) from eastern Quebec. *Mol. Ecol.* **7**, 1655–1662.

Godfray, H. C. J. 2002 Challenges for taxonomy. *Nature* **417**, 17–19.
Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond.* B **270**, 313–322. (DOI 10.1098/rspb.2002.2218.)
Ingaki, Y., Ehara, M., Watanabe, K. I., Hayashi-Ishimaru, Y. & Ohama, T. 1998 Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. *J. Mol. Evol.* **47**, 378–384.
Johns, G. C. & Avise, J. C. 1998 A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome *b* gene. *Mol. Biol. Evol.* **15**, 1481–1490.
Mallet, J. & Willmot, K. 2003 Taxonomy: renaissance or Tower of Babel? *Trends Ecol. Evol.* **18**, 57–59.
Palmer, J. D. 1992 Mitochondrial DNA in plant systematics: applications and limitations. In *Molecular systematics of plants* (ed. P. S. Soltis, D. E. Soltis & J. J. Doyle), pp. 36–49. New York: Chapman & Hall.
Schmidt, T. R., Wu, W., Goodman, M. & Grossman, L. I. 2001 Evolution of nuclear and mitochondrial encoded subunit interaction in cytochrome *c* oxidase. *Mol. Biol. Evol.* **18**, 563–569.
Shearer, T. L., Van Oppen, M. J. H., Romano, S. L. & Worheide, G. 2002 Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Mol. Ecol.* **11**, 2475–2487.
Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74.
Willett, C. S. & Burton, R. S. 2001 Viability of cytochrome *c* genotypes depends on cytoplasmic backgrounds in *Tigriopus californicus*. *Evolution* **55**, 1592–1599.
Wu, W., Schmidt, T. R., Goodman, M. & Grossman, L. I. 2000 Molecular evolution of cytochrome *c* oxidase subunit 1 in primates: is there coevolution between mitochondrial and nuclear genomes. *Mol. Phylogenet. Evol.* **17**, 294–304.

Visit http://www.pubs.royalsoc.ac.uk to see an electronic appendix to this paper.