Progress in Biological Sciences
Vol. 4, Number 2, Summer / Autumn 2014/167-178

Appraisal of the entire mitochondrial genome for DNA barcoding in birds

Received: January 5, 2014; Accepted: March 30, 2014

Leila Nourani[1], Mansour Aliabadian[1,2]*

1. Department of Biology, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran
2. Research Department of Zoological Innovations, Institute of Applied Zoology, Faculty of Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

ABSTRACT

DNA barcoding based on a standardized region of 648 base pairs of mitochondrial DNA sequences from Cytochrome *C* Oxidase 1 (*COX1*) is proposed for animal species identification. Recent studies suggested that DNA barcoding has been effective for identifying 94% of bird species. The proposed threshold of 10 times the average intraspecific variation could be used for the identification and delimitation of new species. As a different part of the mitochondrial DNA evolves at various mutation rates, they show a variety of capabilities to distinguish taxa to species level. In order to compare the efficiency of protein-coding genes (PCGs) in birds, the complete genome of 310 birds, including 12 mitochondrial genes (except *ND6*) and barcoding the region of *COX1,* were examined. We concentrated on the intra- and inter-specific variations and the degree of mutational saturation as criteria for our evaluations. Some genes like *ATP8*, *ND2* and *ND5* showed the greatest divergence in intra- and inter-specific variations. The overlap between intra- and inter-specific variability for all genes is still troublesome. Our results may have been influenced by the sample size because our data were not representative of all bird species. More additional taxa may shed light more on DNA barcoding candidate genes.

Keywords: Aves, *COX1*, interspecific distance, intraspecific distance, mtDNA, phylogeny.

* **Corresponding author:** Aliabadi@um.ac.ir

Introduction

The vertebrate mitochondrial circular genome is a small and extrachoromosomal DNA, which has between 14-20 kb and encodes 37 genes. The mitochondrial genome of animals includes 13 protein coding genes (PCGs), which cooperate in electron transport machinery. Maternal inheritance, limited recombination in some taxonomic groups, lack of introns, higher rate of base substitution and numerous copies of mtDNA are the reasons for using mtDNA more than nuclear DNA (1, 2, 3, 4, 5) for such a study.

To conclude metazoan phylogenetic relationships, scientists use complete mitochondrial genomes more than just one single informative gene to disclose some genome-level factors, such as the rearrangement of genes, which are more important characters for studies of evolutionary relationships (6, 7, 8). The most important subject in the success of DNA barcoding is choosing a reliable barcode candidate (9). DNA barcoding based on the COX1 marker has been successful in various species, but it has been discussed for some invertebrates (10), and a newly diverged group which has a lower rate of diversity (11). The first study barcoding the DNA of 25% of North American bird species showed that the mean inter-specific variation was 20 times greater than intra-specific variation (12), and an obvious space between them was suggested as a standard threshold for delimiting new species. These thresholds for avian species had about 2.7% sequence divergence. Therefore, it is still necessary to search for alternative DNA barcodes to avoid an exclusive reliance on *COX1*.

Given the increasing availability of complete mitochondrial genomes from a range of taxa, marker choice is no longer constrained by the accessibility of universal primers (13). Additionally, nowadays extensive bioinformatics methods are present to estimate the effectiveness of each protein coding genes. Till now, more than 18 papers have been published on the DNA barcoding of bird species (14). Most of them have shown that the individuals have been identified down to species level, for example in Scandinavian birds, with a success range of 94% (15); however, some studies have pointed out the weakness and limits of this approach (16). In this study we examined the effectiveness of the complete mitochondrial genome to test the ability of *COX1* as a DNA barcode candidate gene for the delimitation of bird species. We concentrated on the intra- and inter-specific variations and the degree of mutational saturation as criteria for our evaluations.

Material and Methods

A total of 310 complete mitochondrial genomes (except *Cytb* with 282 taxa) were retrieved from the National Center for Biotechnology Information (NCBI at www.ncbi.nlm.nih.gov/genome) and BOLD system at (www.barcodinglife.org) in June 2013. If the same genome was restored from different databases, duplicate sequences were omitted from our profiles. In order to exclude the ambiguity brought about by any NUMTs that might have entered our dataset, we used all the mitochondria gene sequences available on GenBank. We translated DNA sequences into amino acids and checked for stop codons. We did not observe any NUMTs in our dataset.

The whole genome was partitioned into 13 PCGs profiles and *ND6* was excluded because it has many indels and its place on the light strand of the mitochondrial genome. Sequences

were aligned using the online MAFFT software (www.mafft.cbrc.jp/alignment/server) with default parameter settings and (Bioedit, Tom Hall), also (MEGAv4.0., Koichiro Tamura *et al*.) for all gene profiles were used to reconstruct a phylogenetic tree with the NJ method using Kimura two Parameter (K2P) models (21, 22). The K2P distance is the most effective model when genetic distances are low (2). In order to determine the quantity of sequence divergence, intra- and inter-specific variability was calculated for each gene profile using MEGA v4.0 with the K2P model. Average K2P distances were calculated based on pairwise comparisons of all sequences. The output matrix of pairwise comparison was used later on to compute intra- and inter-specific pairwise distances by a converter program (ExcaliBAR v.1.1, Aliabadian *et al*., Iran) (23).

In order to understand the best-fit evolutionary model of 13 gene profiles, the maximum-likelihood (ML) model of nucleotide evolution was chosen by comparing values of the Akaike Information Criterion (AIC) (ModelTest v.3.7., David Posada and Keith. A. Crandall) (24). Additionally, DNA saturation was estimated to interrogate saturation collections in relation to K2P distance. To compare patterns of DNA saturation among 13 gene profiles, we plotted %Ti value against pairwise K2P distances (25).

Results

In order to estimate the frequency of bird sequences in GenBank, we revised all PCGs till June 2013. Among 13 protein coding-genes, *ND2* and *ND4L* had the most and the least frequency with 16,246 and 71 available sequences, respectively. The number of intra- and inter-specific pairwise comparisons for 13 gene profiles in Table 1 was between 588-620 and 40,177 to 48,835, respectively.

Table 1. Number of intra- and interspecific pairwise comparisons, mean of intra- and interspecific distances i.e. for each 13 gene profiles

Gene	No. Interspecific pairwise comparisons	No. Intraspecific pairwise comparisons	Mean of interspecific distances	Mean of intraspecific distances
ATP6	48835	620	26%	0.23%
ATP8	47307	588	32%	0.05%
COX1	47307	588	18%	0.21%
COX2	47307	588	20.8%	0.21%
COX3	47586	619	19%	0.23%
Cytb	40177	578	20.9%	0.27%
ND1	47307	588	25%	0.18%
ND2	47307	588	29%	0.15%
ND3	47307	588	24%	0.26%
ND4	47617	588	25%	0.22%
ND4L	47307	588	23%	0.32%
ND5	47307	588	27.2%	0.12%
Barcoding	47307	588	19.3%	0.28%

The total number of species was the same in all profiles (310 taxa) except *Cytb* with 282 taxa. We had 186 species of 134 genera and *Cytb* was an exception with 172 species of 123 genera. Average K2P distances were estimated for each gene profile to show whether there is any sequences variability among different parts of the mitochondrial genome. The average intraspecific distances range were from 0.05% (*ATP8*) to 0.32% (*ND4L*) and the average interspecific distances range in Table 1 were from 18% (*COX1*) to 32% (*ATP8*). Among all gene profiles *ATP8*, *ND2* and *ND5* in Figure 1 showed the greatest variation between intra- and inter-specific distances. The *ND2* gene was distinct in having a smaller proportion of invariable sites (21.54%) and less heterogeneity of substitution rates was observed in *COX3* (shape parameter = 0.3904), as shown in Table 2 and Figure 2.

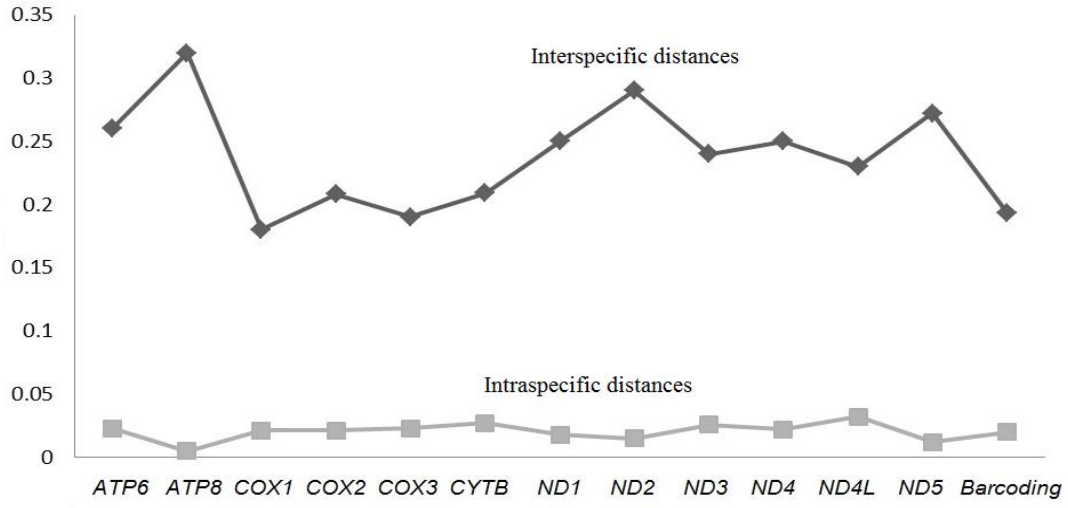


Figure 1. Mean intra- and interspecific distances for 13 gene profiles. Black markers show interspecific and gray markers show intraspecific distances. The gene profiles are placed on the x-axis and intra- and interspecific distances on y-axis.

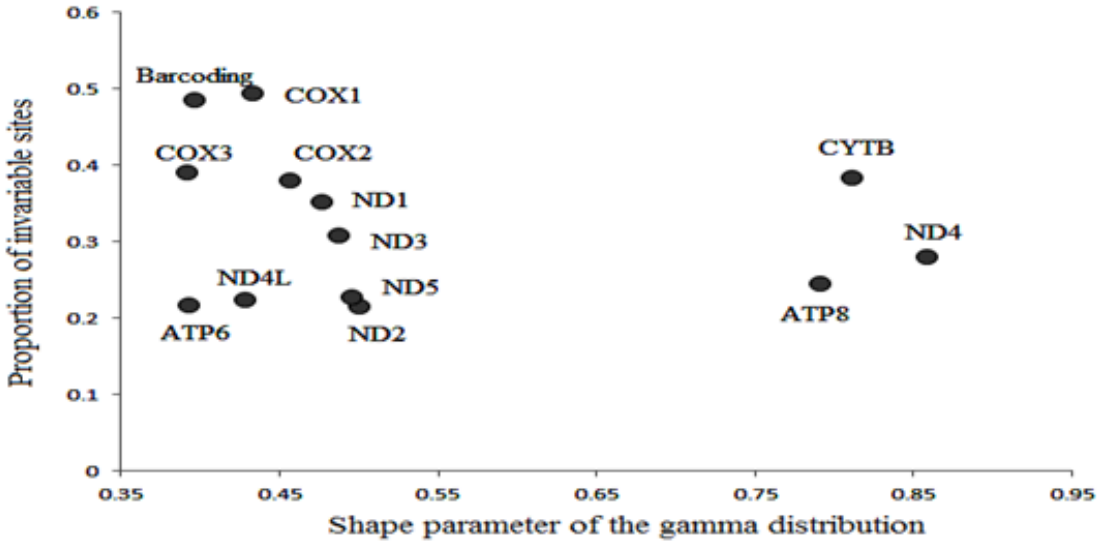


Figure 2. Evaluated values of substitution model parameter for 13 gene profiles. The shape parameter of the gamma distribution and proportion of invariable sites shows on the x and y- axis respectively.

Table 2. Number of conserved sites, variable sites, parsimony sites, singleton sites, and frequency of each nucleotide for all 13 mitochondrial gene profiles

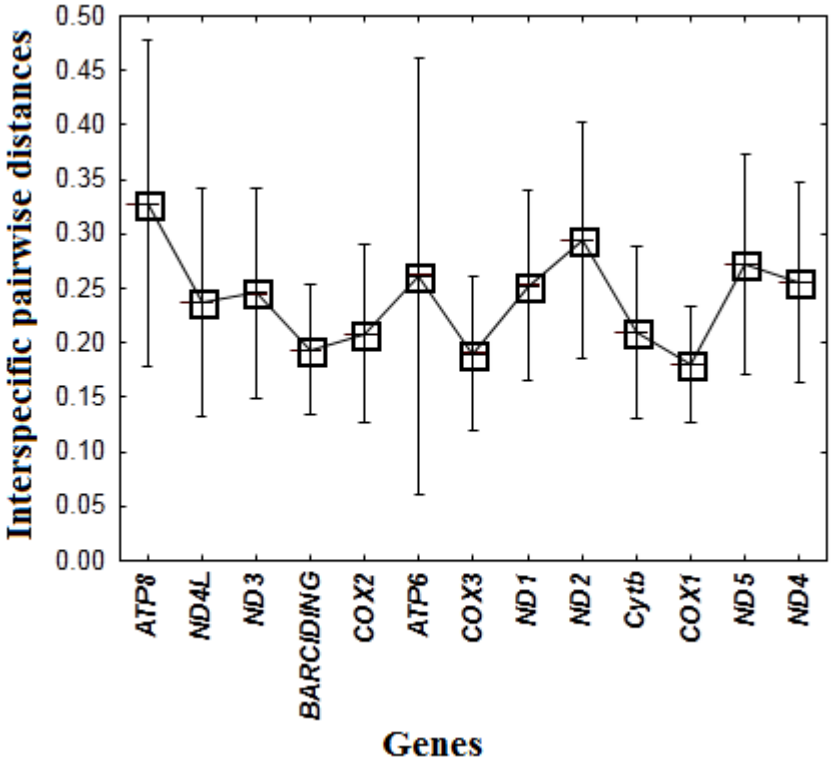
Gene	Base pair	Conserved site	Variable site	Parsimony site	Singleton site	T	C	A	G
ATP6	684	166	518	499 (%72.95)	19 (%2.77)	24.6	36.1	29.4	10
ATP8	165	42	123	120 (%72.72)	3 (%1.81)	23.6	36.5	34.2	5.8
Barcoding	648	337	311	295 (%45.52)	16 (%2.46)	25.2	32.5	25.6	16.6
COX1	1544	801	743	703 (%45.52)	40 (%2.59)	25.6	30.9	27.3	16.3
COX2	684	278	406	392 (%57.3)	14 (%2.04)	23.8	31.4	30.1	14.7
COX3	784	333	451	426 (%54.33)	25 (%3.18)	24.4	32.5	27.8	15.3
Cytb	1143	451	692	642 (%56.16)	50 (%4.37)	24.7	34.9	27.9	12.5
ND1	967	362	605	580 (%59.97)	25 (%2.58)	26	33.7	27.1	13.1
ND2	1041	254	787	746 (%71.66)	41 (%3.93)	23.7	34.7	31.7	9.9
ND3	352	119	233	222 (%63.06)	11 (%3.12)	25.9	33.5	28.2	12.4
ND4	1378	399	979	910 (%66.03)	69 (%5)	23.5	35.4	30.6	10.5
ND4L	297	77	220	202 (%68.01)	18 (%6.06)	25	34.3	27.6	13.1
ND5	1816	469	1347	1286 (%70.81)	61 (%3.35)	23	34.3	31.3	11.4

The number of conserved sites, variable sites, parsimony sites and singleton sites are listed. Furthermore, the frequency of nucleotide bases is mentioned in Table 2. *ATP6* (24.26%) had the lowest rate of conserved sites and the highest proportion of variable sites (75.74%) and the barcoding region of *COX1* (52%) had the greatest value for conserved sites and the lowest rate for variable sites (48%).

Discussion

Our results based on the NJ trees and divergence between intra- and inter-specific distances showed that the barcoding region of *COX1* is not the only suitable marker as a barcode for our samples. Some genes like *ATP8* (165 bp), *ND2* (1041 bp) and *ND5* (1816 bp) showed the greatest divergences in intra- and inter-specific distances between 13 gene profiles. In a similar study on eutherian mammals, some genes like *ND5*, *Cytb* and *ATP8* showed the greatest divergence and *COX1* had fewer divergences in comparison to other genes (26). One important subject with the use of mitochondrial DNA is due to NUMTs, which are responsible for the misidentification in mitochondrial studies and phylogenetic reconstruction (27, 28).

Studying the Mean ± 0.95 confidence interval of intra- and inter-specific pairwise distances for 13 mitochondrial gene profiles shown in Figures 3A and 3B did not show any meaningful relation between gene nucleotide base pairs and intra- and inter-specific pairwise comparisons. In another approach to increasing the number of base pairs in gene nucleotides we did not observe any specific relation in intra- and inter-specific pairwise comparisons.



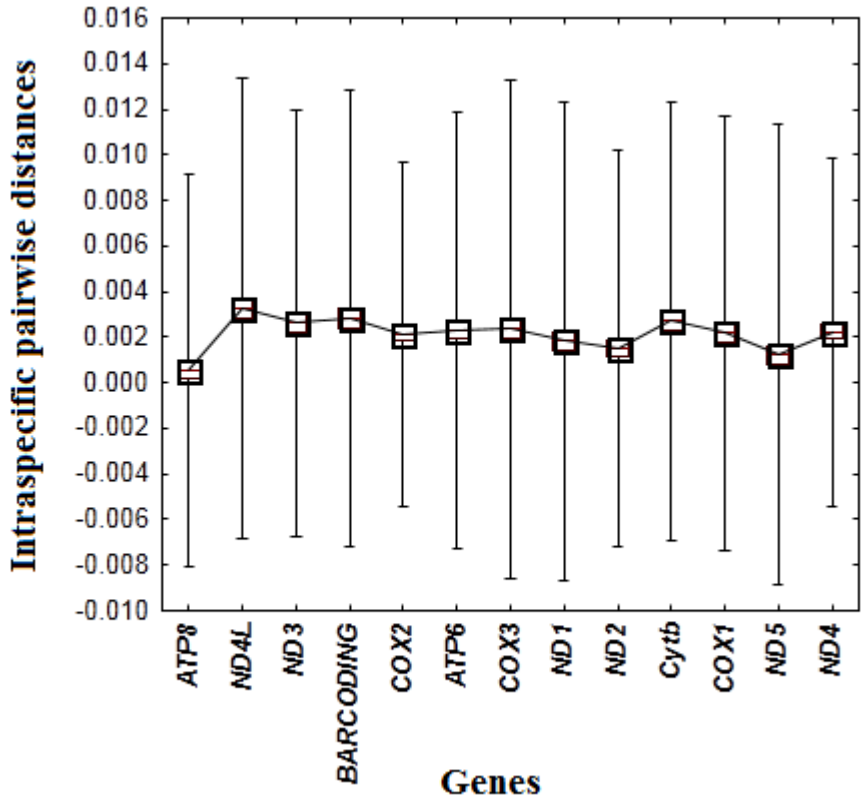


Figure 3. Interspecific (A) and intraspecific (B) pairwise distances of 13 mitochondrial gene profiles. Boxes show 0.95 confidence interval, and bars show +/- 2 SD.

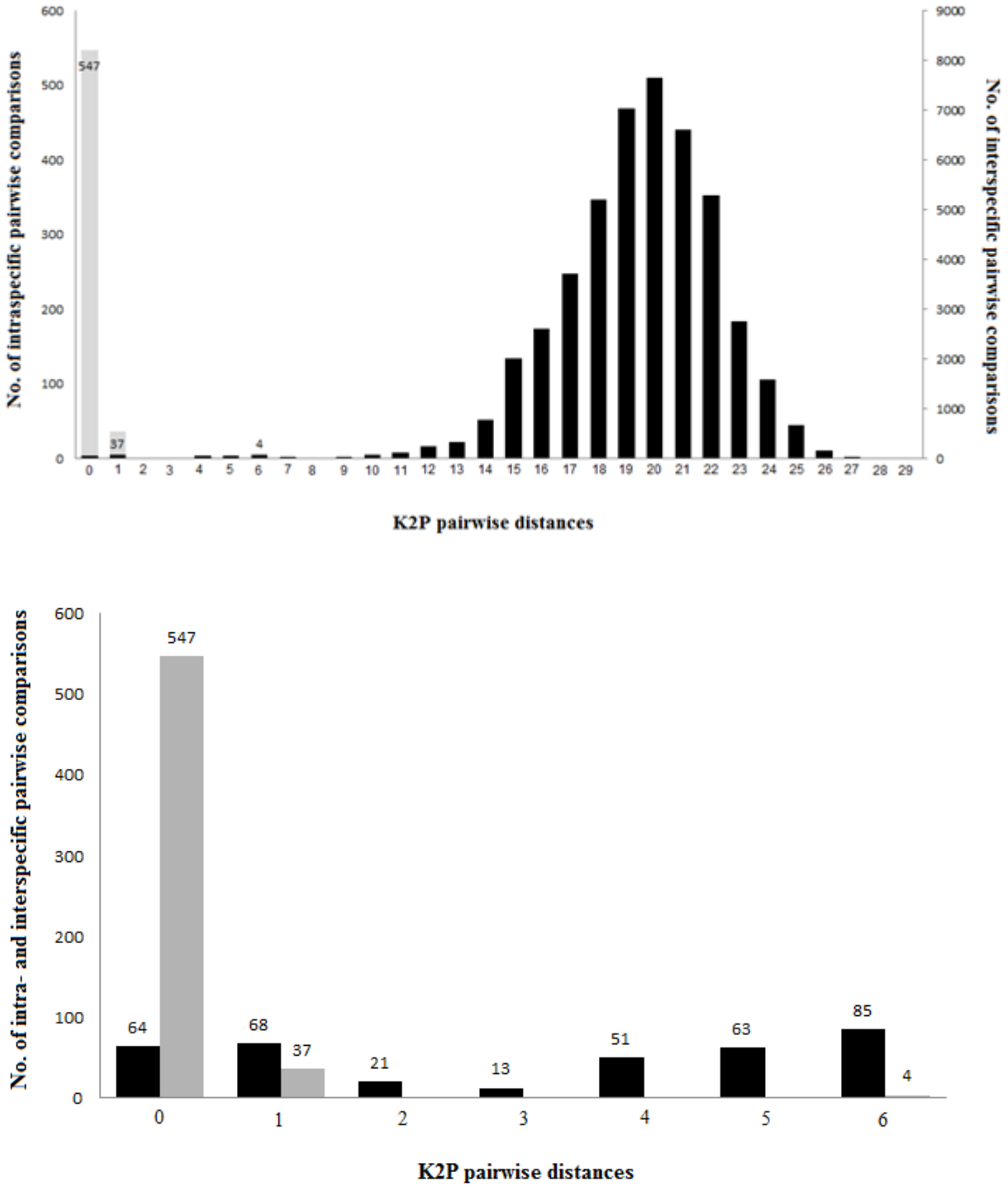


Figure 4. Comparison of intraspecific and interspecific pairwise K2P distances. A) Intra- and interspecific pairwise comparisons for barcoding region of *COX1*, B) Re-presentation of data in A to show degree of overlap between intraspecific and interspecific pairwise K2P distances of barcoding region of *COX1*. Grey bars, intraspecific comparisons; black bars, interspecific comparisons.

Among eukaryotes, mitochondrial protein coding genes are highly conserved with putative phylogenetic information, but *ND4L* and *ATP8* show less conserved sites (29). In

our results the range of conserved sites was between 24.26% (*ATP6*) and 52% (barcoding region of *COX1*). Our results in Figures 4A and 4B showed that the *COX1* barcoding region is a suitable marker in mitochondrial protein coding genes as a universal DNA barcode. The mean of intra- and inter-specific distances for the barcoding region of *COX1* was 0.28% and 19.3%, respectively, as shown in Table 1 and Figure 1.

The overlap between intra- and inter-specific variability in all our gene profiles was an important problem. The misnamed or misidentified taxa in GenBank or the existence of ring species might be the reason for the lower range of intra- and inter-specific distances (30) and the highest intraspecific distances are because of mistakenly identified samples retrieved from GenBank or the BOLD system, for example.

Table 3. False negative values for barcoding region of *COX1* gene

Species	Species	relationship	K2P distances (%)
Syrmaticus ellioti	*Syrmaticus humiae*	Hybrid	0%
Gallus gallus	*Gallus sonneratii*	Hybrid	0%
Calyptorhynchus latirostris	*Calyptorhynchus baudinii*	Hybrid	0.15%
Gallus gallus	*Gallus sonneratii*	Hybrid	0.31%
Gallus gallus	*Gallus sonneratii*	Hybrid	0.47%
Pycnonotus taivanus	*Pycnonotus sinensis*	Hybrid	0.62%
Gallus gallus	*Gallus sonneratii*	Hybrid	0.62%
Pycnonotus sinensis	*Pycnonotus taivanus*	Hybrid	0.62%
Thalassarche melanophrys	*Diomedea chrysostoma*	Hybrid	0.93%
Gallus gallus	*Gallus sonneratii*	Hybrid	0.94%
Gallus gallus	*Gallus varius*	Hybrid	1.25%
Grus nigricollis	*Grus monacha*	Sister taxa	1.25%
Gallus varius	*Gallus gallus*	Hybrid	1.41%
Grus monacha	*Grus grus*	hybrid	1.72%
Apteryx owenii	*Apteryx haastii*	Sister taxa	1.72%
Grus nigricollis	*Grus grus*	Hybrid	1.72%
Platalea leucorodia	*Platalea minor*	Sister taxa	1.72%
Gallus gallus	*Gallus varius*	Hybrid	1.73%
Ciconia boyciana	*Ciconia ciconia*	Hybrid	2.21%
Anser anser	*Anser albifrons*	Hybrid	2.21%
Grus antigone	*Grus rubicunda*	Hybrid	2.36%
Grus virgo	*Grus paradisea*	Hybrid	2.37%
Grus nigricollis	*Grus americana*	Sister taxa	2.53%
Grus monacha	*Grus americana*	Sister taxa	2.53%
Grus grus	*Grus americana*	Hybrid	2.69%

As the proposed threshold of 10 times the average intraspecific variation could be used for the identification and delimitation of new species, in our study, we found the same threshold (2.8%) for species discrimination. We observed four pairs of *Gallus sonnerati*, which showed a higher value than the proposed threshold (5.68%). The species with a value less than 2.8% in interspecific pairwise comparisons ranged from 0% to 2.69%. According to recent studies, *COX1* as a universal barcode was not successful in identifying 48% of parapatric hybrid species (31) and our results in Table 3 showed that from all comparisons pairs 10.59% of comparisons belong to hybrid taxa and 3.12% belong to sister taxa. The barcoding region of *COX1* was not successful in distinguishing them.

In all gene profiles, the percentage of variable sites was more than conserved sites but in the *COX1* gene and the barcoding region of *COX1* showed the opposite relationship. The best-fit evolutionary model was GTR+I+G for each gene profile, using the K2P model, which underestimates the frequency of multiple substitution in variable nucleotide sites (32). We found that the high percentage of %Ti disappeared when nucleotide distances increased and this plot was similar for all gene profiles. We plotted nucleotide distance using the K2P model against %Ti value in Figure 5. The low value of %Ti might be associated with the fact that compared with transversions and transitions would be less evident if K2P did not correct for the multiple hits hidden behind transversions for closely related pairs (26).

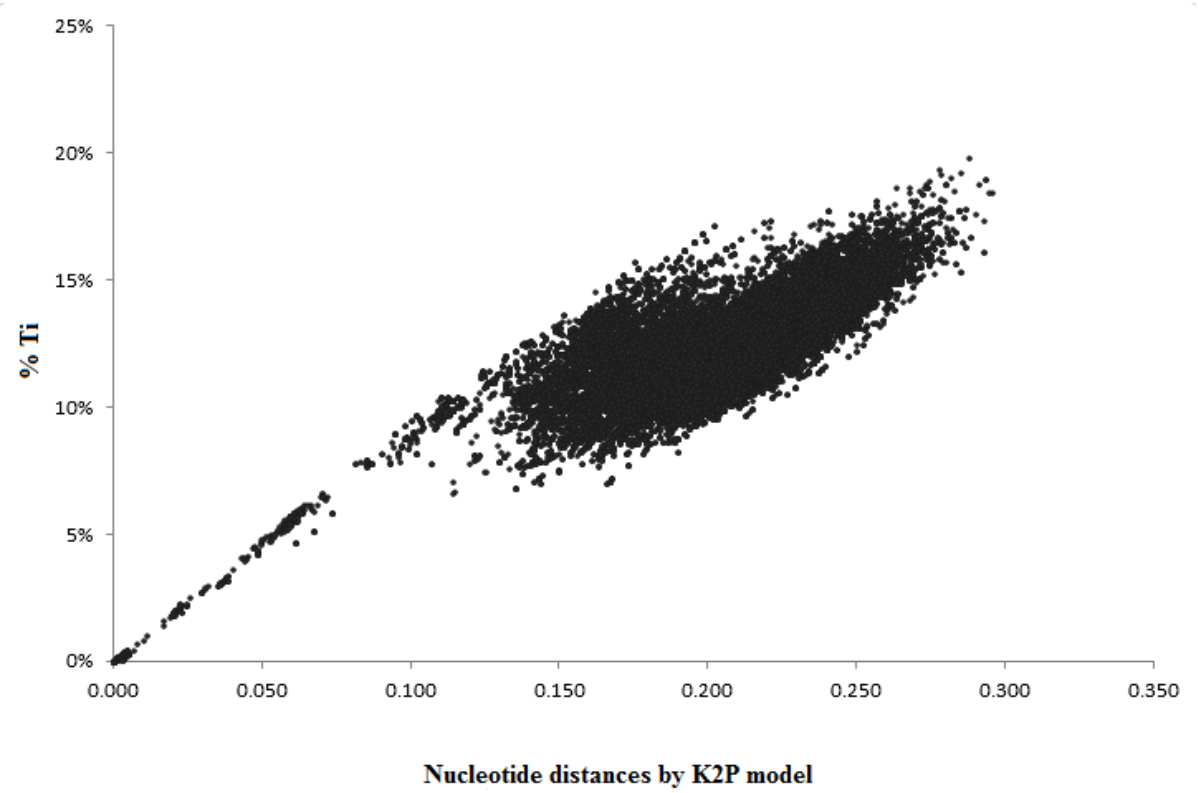


Figure 5. Nucleotide distance using K2P model against %Ti value

Conclusion

Our estimation of examine effectiveness of DNA barcoding indicated that other protein coding genes can be used as a complementary barcode in addition to the *COX1,* currently used as a universal barcode to study the phylogenetic relation within bird groups. Although some genes like *ATP8* with a high range of saturation and short length of nucleotides cannot be useful for eutherian mammals (26), these genes show the greatest divergence between intra- and inter-specific distances in bird mitochondrial gene profiles. In comparison to the single gene datasets, complete mitochondrial sequences can provide better phylogenetic resolution to show the evolutionary relationship (33) and each partitioned gene profile from the complete mitochondrial genome with different evolutionary models may increase the preciseness and support of deep phylogenetic relationships (34). The overlap between intra- and inter-specific variability for all genes is still troublesome. Our results may have been influenced by the sample size because our data was not representative of all bird species. More additional taxa may shed light on more DNA barcoding candidate genes.

Acknowledgements

We would like to thank those who prepared and submitted mitochondrial genome sequences of birds and who provided helpful comments on an earlier version of this manuscript. This work was partially funded by a grant from the Ferdowsi University of Mashhad to L.N (3/17427).

.

REFERENCES

1. Boore, J.L. (1999) Survey and Summery, Animal mitochondrial genomes. *Nucleic. Acids. Res.,* 27, 1767-1780.
2. Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B. Biol. Sci*., 270, 313-321.
3. Rokas, A., Ladoukakis, E. and Zouros, E. (2003) Animal mitochondrial DNA recombination revisited. *Trends. Ecol. Evol.,* 18, 411-417.
4. Ladoukakis, E.D. and Zouros, E. (2001) Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis*) mitochondrial DNA. *Mol. Biol. Evol.,* 18, 1168-1175.
5. Hoarau, G., Holla, S., Lescasse, R., Stam, W.T. and Olsen, J.L. (2002) Heteroplasmy and evidence for recombination in the mitochondrial control region of the lfatifsh *Platichthys flesus*. *Mol. Biol. Evol.,* 19, 2261-2264.
6. Boore, J.L. and Brown, W.M. (1998) Big trees from little genomes, mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev*., 8, 668-674.
7. Boore, J.L., Lavrov, D.V. and Brown, W.M. (1998) Gene translocation link s insects and crustaceans. *Nature*., 392, 667-668.
8. Boore, J.L., Macey, J.R. and Medina, M. (2005) Sequencing and comparing whole mitochondrial genomes of animals. *Methods. Enzymol*., 395, 311-348.
9. Waugh, J. (2007) DNA barcoding in animal species, progress, potential and pitfalls. *Bioessays*., 29, 188-197.
10. Wiemers, M. and Fiedler, K. (2007) Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). *Front. Zool.,* 4, 8.
11. Moritz, C. and Cicero, C. (2004) DNA barcoding: promise and pitfalls. *PloS. Biol.,* 2, 1529-1531.
12. Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S. and Francis, C.M. (2004) Identification of birds through DNA barcodes. *PLoS. Biol*. 2, e312.
13. Miller, A.D. and Austin, C.M. (2006) The complete mitochondrial genome of the mantid shrimp *Harpiosquilla harpax*, and a phylogenetic investigation of the Decapoda using mitochondrial sequences. *Mol. Phylogenet. Evol*., 38, 565-574.
14. Taylor, H.R. and Harris, W.E. (2012) An emergent science on the brink of irrelevance, a review of the past 8 years of DNA barcoding. *Mol. Ecol. Res.*, 12, 377-388.
15. Johnsen, A., Rindal, E., Ericson, P.G.P., Zuccon, D., Kerr, K.C.R., Stoeckle, M.Y. and Lifjeld, J.T. (2010) DNA barcoding of Scandinavian birds reveals divergent lineages in trans-Atlantic species. *J. Ornith*., 151, 565-578.
16. Marshall, E. (2005) Will DNA barcodes breathe life into classification?. *Science*., 307, 1037.
17. Nucleotide Home of NCBI (http://www.ncbi.nlm.nih.gov/nuccore).
18. BOLD system (http://www.barcodinglife.org).

19. MAFFT software online (http://www.mafft.cbrc.jp/alignment/server).

20. Hall, T.A. (1999) BioEdit, a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic. Acids. Symp. Ser*., 41, 95-98.

21. 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.

22. Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*., 16, 111-120.

23. Aliabadian, M., Nijman, V., Mahmoudi, A., Naderi, M., Vonk, R. and Vences, M. (2014) ExcaliBAR, a simple and fast software utility to calculate intra- and interspecific distances from DNA barcodes, Contributions to Zoology., 83, 79-83.

24. Posada, D. and Crandall, K.A. (1998) MODELTEST, testing the model of DNA substitution. *Bioinformatics*., 14, 817-818.

25. Roe, A.D. ad Sperling, F.A. (2007) Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Mol. Phylogenet. Evol*., 44, 325-345.

26. Luo, A., Zhang, A., Ho, S.Y.W., Xu, W., Zhang, Y., Shi, W., Cameron, S.L., and Zhu, C. (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding, a case study using eutherian mammals. *BMC. Genomics*., 12, 84. doi:10.1186 /1471-2164-12- 84.

27. Richly, A. and Leister, D. (2004) NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.,* 21, 1081-1084.

28. Song, H., Buhay, J.E., Whiting, M.F. and Crandall, K.A. (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.

29. Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol*., 17, 540-552.

30. Crochet, P.A., Lebreton, J.D. and Bonhomme, F. (2002) Systematics of large white-headed gulls: patterns of mitochondrial DNA variation in western European taxa. *Auk*., 119, 603-620.

31. Aliabadian, M., Kaboli, M., Nijman, V. and Vences, M. (2009) Molecular identification of birds: performance of distance-based DNA barcoding in three genes to delimit parapatric species. *PLoS One*., 4, e4119.

32. Nei, M. and Kumar, S. (2000) Molecular Evolution and Phylogenetics, New York, Oxford University Press.

33. Zardoya, R. and Meyer, A. (2001) On the origin of and phylogenetic relationships among living amphibians. *PNAS*., 98, 7380-7383.

34. Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P. and Nieves-Aldrey, J.L. (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol*., 53, 47–67.