Bayesian, Maximum Parsimony and UPGMA Models for Inferring the Phylogenies of Antelopes Using Mitochondrial Markers

Haseeb A. Khan, Ibrahim A. Arif, Ali H. Bahkali, Ahmad H. Al Farhan and Ali A. Al Homaidan

Molecular Fingerprinting and Biodiversity Unit, Prince Sultan Research Chair Program in Environment and Wildlife, College of Science, King Saud University, Riyadh, Saudi Arabia.

Abstract: This investigation was aimed to compare the inference of antelope phylogenies resulting from the 16S rRNA, cytochrome-b (cyt-b) and d-loop segments of mitochondrial DNA using three different computational models including Bayesian (BA), maximum parsimony (MP) and unweighted pair group method with arithmetic mean (UPGMA). The respective nucleotide sequences of three Oryx species (*Oryx leucoryx*, *Oryx dammah* and *Oryx gazella*) and an out-group (*Addax nasomaculatus*) were aligned and subjected to BA, MP and UPGMA models for comparing the topologies of respective phylogenetic trees. The 16S rRNA region possessed the highest frequency of conserved sequences (97.65%) followed by cyt-b (94.22%) and d-loop (87.29%). There were few transitions (2.35%) and none transversions in 16S rRNA as compared to cyt-b (5.61% transitions and 0.17% transversions) and d-loop (11.57% transitions and 1.14% transversions) while comparing the four taxa. All the three mitochondrial segments clearly differentiated the genus Addax from Oryx using the BA or UPGMA models. The topologies of all the gamma-corrected Bayesian trees were identical irrespective of the marker type. The UPGMA trees resulting from 16S rRNA and d-loop sequences were also identical (*Oryx dammah* grouped with *Oryx leucoryx*) to Bayesian trees except that the UPGMA tree based on cyt-b showed a slightly different phylogeny (*Oryx dammah* grouped with *Oryx gazella*) with a low bootstrap support. However, the MP model failed to differentiate the genus Addax from Oryx. These findings demonstrate the efficiency and robustness of BA and UPGMA methods for phylogenetic analysis of antelopes using mitochondrial markers.

Keywords: antelopes, mitochondrial DNA, phylogenetic trees, bioinformatics, Bayesian, maximum parsimony, UPGMA

Introduction

The antelope Arabian Oryx was extirpated from the wild as a result of massive hunting during early 1970s (Henderson, 1974). Fortunately, the efforts of captive breeding programs succeeded to preserve the Arabian Oryx, which was later reintroduced in certain protected areas (Spalton et al. 1999; Ostrowski et al. 1998; Mesochina et al. 2003). Recently, Iyengar et al. (2007) have recommended maintaining a global perspective for the captive genetic management of Oryx. Individuals from various management programs and regions need to be effectively utilized for sustained future captive breeding in order to ensure that the vital remnants of genetic diversity are retained and represented in future reintroduction programs (Iyengar et al. 2007). It has been suggested that molecular methods can significantly contribute to the captive breeding and reintroduction strategies for conservation of various endangered animals such as Oryx antelopes (Russello and Amato, 2007).

Molecular fingerprinting based on the nucleotide sequence analysis of various mitochondrial genes plays an important role in studying evolutionary relationship among various taxa. Besides its exclusive maternal inheritance and lack of recombination, different segments of mitochondrial DNA (mtDNA) also possess unique features such as conservativeness in protein-coding regions and high variability in non-coding sequences (Ingman et al. 2000; Olivo et al. 1983). Consequently, the evolutionary rate of mtDNA tends to be variable for different regions and has been utilized to examine various levels of phylogenetic relationships. The 12S rRNA gene sequences being highly conserved, are applied to illustrate higher levels of phylogenies (phyla or subphyla) whereas the 16s rRNA sequences are mainly used for phylogenetic studies at mid-categorical levels (families or genera) (Gerber et al. 2001). Since the mitochondrial protein-coding

Correspondence: Haseeb A. Khan Ph.D., MRACI (Aus.), FRSC (U.K.), Chair Professor, Saudi Biological Society, College of Science, Bld 5, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia. Email: khan_haseeb@yahoo.com

genes and the d-loop evolve comparatively faster they are considered as powerful tools for inferring evolutionary history in mid to lower categorical levels such as genera and species.

Probabilistic modeling of sequence evolution has now become inevitable in phylogenetic inference (Felsenstein, 2001). Despite a positive impact of statistical revolution, the emergence of sophisticated evolutionary models has placed a burden on researchers to select the model most appropriate for their data. It is intriguing that the bioinformatics tool used for phylogenetic analysis may have some influence on the topologies of the resulting trees. An inappropriate choice of evolutionary model can affect the outcome of any phylogenetic analysis by incorrectly estimating tree topologies (Penny et al. 1994; Bruno and Halpern, 1999). Bayesian (BA), maximum likelihood (ML) or unweighted pair group method with arithmetic mean (UPGMA) and maximum parsimony (MP) are the main phylogenetic approaches that are often used side by side. While the choice between them has been contentious at times, they frequently give similar results and if they don't, they can complement each other (Liberles, 2005). In this investigation, we have compared BA, MP and UPGMA methods for phylogenetic analysis of Oryx antelopes using 16S rRNA, cytochrome-b (cyt-b) and d-loop sequences of mtDNA.

Methods

The sequences of 16S rRNA, cyt-b and d-loop of the three Oryx species including Arabian Oryx (*Oryx leucoryx*), Scimitar Horned Oryx (*Oryx dammah*) and Plains Oryx (*Oryx gazella*) were obtained from GenBank. The respective sequences of Addax (*Addax nasomaculatus*) were used as outgroup due to its close relationship to Oryx yet representing a separate sister taxa (Hassanin and Douzery, 1999; Iyengar et al. 2006). The GeneBank accession numbers and the number of nucleotides for the partial sequences of 16S rRNA, cyt-b and d-loop of the four taxa are: *Oryx leucoryx* (U87021, 342; AF036286, 1143; AJ235326, 1253), *Oryx dammah* (U87020, 342; AJ222685, 1143; AJ235324, 1261), *Oryx gazella* (U87022, 342; AF249973, 1140; AJ235325, 1237) and *Addax nasomaculatus* (U87023, 342; AF034722, 1143; AJ235310, 1324) respectively.

Tajima test statistics (Tajima, 1989) and the test of homogeneity of substitution patterns between sequences were performed after sequence alignments, using MEGA4 software (Tamura et al. 2007) while all the positions containing gaps and missing data were eliminated from the dataset (complete deletion option). The probability of rejecting the null hypothesis that sequences have evolved with the same pattern of substitution was judged from the extent of differences in the base composition biases between sequences (disparity index test) whereas a Monte Carlo test (1000 replicates) was used to estimate the respective *P*-values (Kumar and Gadagkar, 2001).

The sequence data were subjected to three different methods of phylogenetic reconstruction: (i) Bayesian (BA), (ii) unweighted pair group method with arithmetic mean (UPGMA) and (iii) maximum parsimony (MP). The gamma-corrected Bayesian inference of phylogeny was conduced using MrBayes software (Huelsenbeck and Ronquist, 2001) and the Bayesian trees were visualized with TreeView software (Page, 1996). For UPGMA method, the phylogenetic analyses were performed using the evolutionary distances computed by maximum composite likelihood method (Sneath and Sokal, 1973; Tamura et al. 2004). For MP method, the maximum parsimonious trees were obtained using the close-neighbor-interchange algorithm in which the initial trees were obtained with the random addition of sequences for 10 replicates (Eck and Davhoff, 1966; Nei and Kumar, 2000). Both UPGMA and MP analyses were performed using MEGA4 software and the bootstrap consensus trees inferred from 1000 replicates were taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985; Tamura et al. 2007).

Results

Both 16S rRNA and cyt-b sequences were perfectly aligned without any insertions/deletions (indels) whereas numerous indels at various sites of different taxa were required to align the sequences of d-loop (please refer to electronic supplementary file). The average frequencies of identical (conserved) sequences between the taxa were 97.65% for 16S rRNA, 94.22% for cyt-b and 87.29% for d-loop (Fig. 1). On an average there were few transitions (2.35%) and none transversions in 16S rRNA as compared to cyt-b (5.61% transitions and 0.17% transversions) and d-loop (11.57% transitions and 1.14% transversions) (Fig. 1).

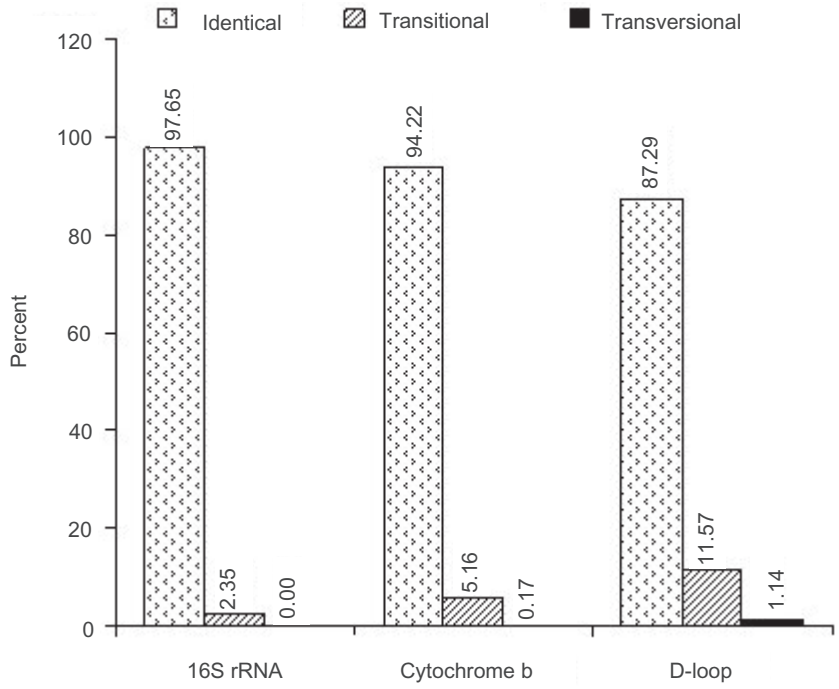

Figure 1. Average frequencies of identical (conserved) and substituted (transitional and transversional) sites observed in sequence comparison for various segments of mtDNA.

The results of Tajima's neutrality are given in Table 1. Both the number of segregating sites (S) and nucleotide diversities (π) were directly correlated and were in the order of 16S rRNA ($S = 17, \pi = 0.025$) < cyt-b ($S = 125, \pi = 0.058$) < d-loop ($S = 270, \pi = 0.122$) (Table 1). The test of homogeneity of substitution patterns showed certain identities and certain variations in disparity index as well as Monte Carlo probability for different mitochondrial markers (Table 2).

The topologies of all the Bayesian trees were identical irrespective of the marker type, which clearly differentiated the genus Addax from Oryx, and grouped *Oryx dammah* with *Oryx leucoryx* (Fig. 2). The UPGMA trees resulting from 16S rRNA and d-loop sequences were also identical (*Oryx dammah* grouped with *Oryx leucoryx*) to Bayesian trees except that the UPGMA tree based on cyt-b showed a slightly different phylogeny (*Oryx dammah* grouped with *Oryx gazella*) with a low bootstrap support (Fig. 3). The MP method failed to differentiate the genus Addax from Oryx and *Addax nasomaculatus* was either grouped with *Oryx leucoryx* (16S rRNA or cyt-b) or with *Oryx gazella* (d-loop) (Fig. 4).

Discussion

In conservation genetics, knowledge of the relatedness between individuals is particularly important for captive breeding programs to recover small populations (Frankham et al. 2002; Montgomery et al. 1997). Genetically impoverished endangered populations often fail to exhibit signs of recovery until crossed with individuals from other populations (Land and Lacy, 2000; Westemeier et al. 1998).

Table 1. Tajima's neutrality test for 4 taxa using different mitochondrial markers.

	Number of sites (m)	Number of segregating sites (S)	Ps = S/m	Nucleotide diversity (π)	Tajima test statistics (D)
16S rRNA	4	17	0.049708	0.025341	–0.667112
Cyt-b	4	125	0.109457	0.058085	–0.283933
D-loop	4	270	0.224439	0.122333	–0.007559

Table 2. The test of homogeneity of substitution patterns for different mitochondrial markers.

	Addax	Oryx leucoryx	Oryx dammah	Oryx gazella
16S rRNA				
Addax	–	0.000	0.000	0.000
Oryx leucoryx	1.000	–	0.018	0.000
Oryx dammah	1.000	0.074	–	0.041
Oryx gazella	1.000	1.000	0.012*	–
Cyt-b				
Addax	–	0.000	0.000	0.000
Oryx leucoryx	1.000	–	0.000	0.000
Oryx dammah	1.000	1.000	–	0.000
Oryx gazella	1.000	1.000	1.000	–
D-loop				
Addax	–	0.053	0.000	0.114
Oryx leucoryx	0.261	–	0.000	0.000
Oryx dammah	1.000	1.000	–	0.000
Oryx gazella	0.124	1.000	1.000	–

The estimates of the disparity index per site are shown for each sequence pair above the diagonal. The *P* values based on Monte Carlo test (1000 replicates) are shown below the diagonal. *$P < 0.05$, statistically significant.

However, if the strategy is to maintain the genetic identity of the population, the introduced individuals should be closely related to the recipient population. Recently, Masembe et al. (2006) have recommended the need of conservation efforts to preserve genetic identity of various oryx groups. Molecular methods play an important role in estimating the relatedness between individuals by comparing the genotypes at a number of informative loci (Sunnucks, 2000). The high mutation rate of mtDNA compared to nuclear genes renders mtDNA sequences to possess high levels of informative variation that could be utilized for resolving taxonomic relationship in conservation genetics using appropriate bioinformatics tools.

We observed no indels in 16S rRNA and cyt-b genes whereas numerous indels were noticed in the aligned sequences of d-loop which is in agreement with an earlier study reporting specific indels in the d-loop of Oryx species (Iyengar et al. 2006). The frequency of conserved sequences was highest in 16S rRNA gene followed by cyt-b and was lowest in d-loop region whereas the converse was true for the substitutions (Fig. 1). Most of the substitutions in the mitochondrial regions studied were transitional indicating a recent species history. Factually, transitions are typically observed more often than transversions in the evolution of real sequences.

The BA model with gamma correction appears to be the most efficient method as it produced identical trees using the nucleotide sequences of any of the three segments of mtDNA (Fig. 2). Bayesian inference has been successfully applied to inference of phylogenetic trees using mitochondrial and nuclear genes (Doudy et al. 2003; Xiong et al. 2002; Ragan et al. 2003). Although likelihood-based approaches have proven to be especially powerful for inferring phylogenetic trees they tend to be prohibitively slow due to the requirement of multidimensional space for possible outcomes (optimal trees) and the computational complexity of bootstrap repetitions. On the other hand, BA phylogenetic inference holds promise as an alternative to ML, particularly for large molecular-sequence data sets. Moreover, BA phylogenetic inference has been shown to be as or more robust to ML, particularly when among-sites rate variation is modeled using a gamma distribution (Mar et al. 2005).

The UPGMA model also produced similar phylogenies to BA model for 16S rRNA and d-loop sequences however cyt-b inferred a different phylogeny (Fig. 3). These differential phylogenies may be associated with comparatively high variations

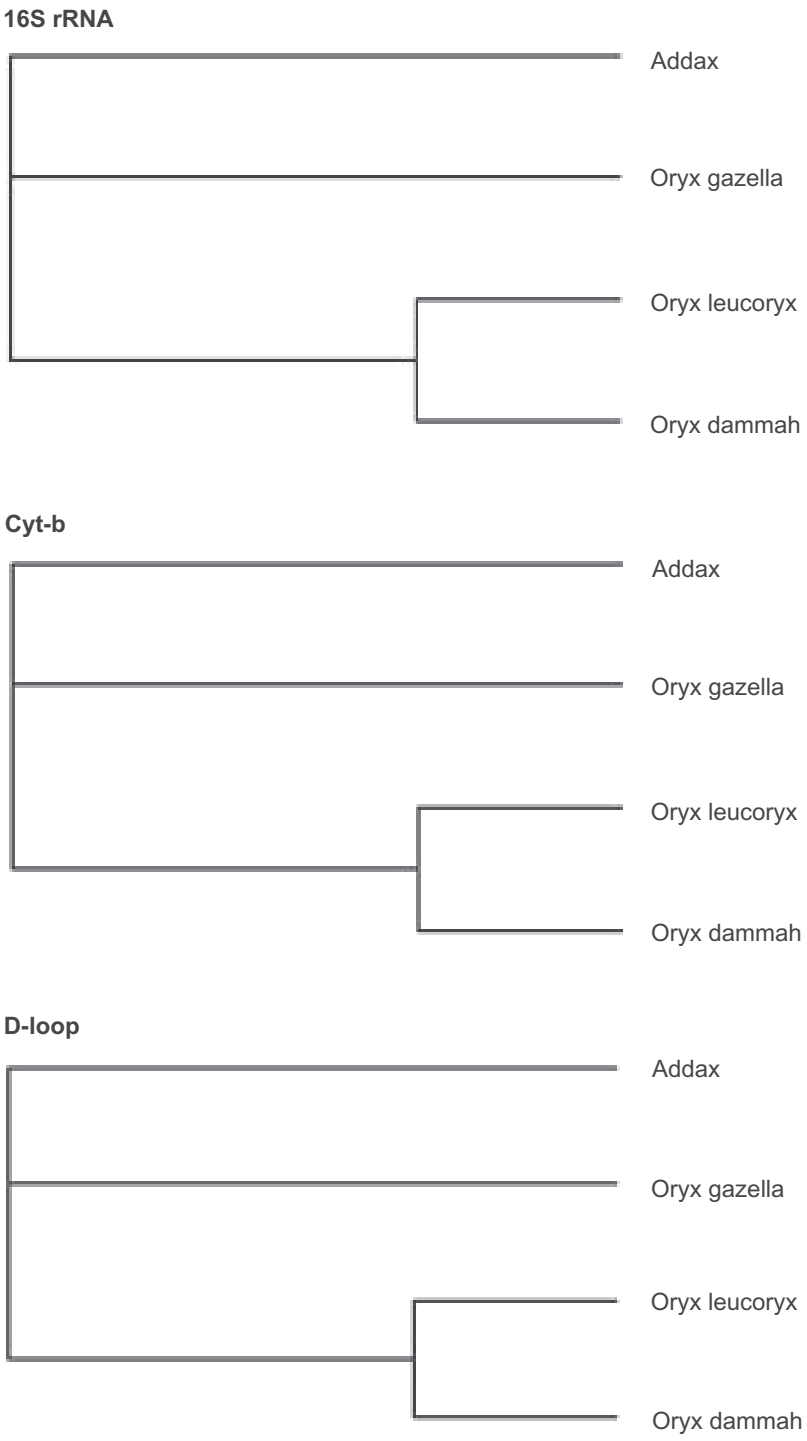

Figure 2. Bayesian method for inferring phylogenetic relationship among various Oryx species using Addax as an outgroup.

in non-coding d-loop than coding cyt-b due to reduced functional constraints and relaxed selection pressure. Although, increased polymorphism in d-loop segment may render it superior to cyt-b for species or sub-species level identification, the possibility of reduced phylogenic information due to back mutations and parallel substitutions in rapidly-evolving d-loop may not be ruled out. It is also important to mention that changing the outgroup species or the length of d-loop segment can significantly alter the topology of phylogenic trees (Iyenger et al. 2006).

The MP model resulted different phylogenetic inferences than those from the BA and UPGMA models (Fig. 4). A certain degree of contradictive phylogeny using mitochondrial markers has been

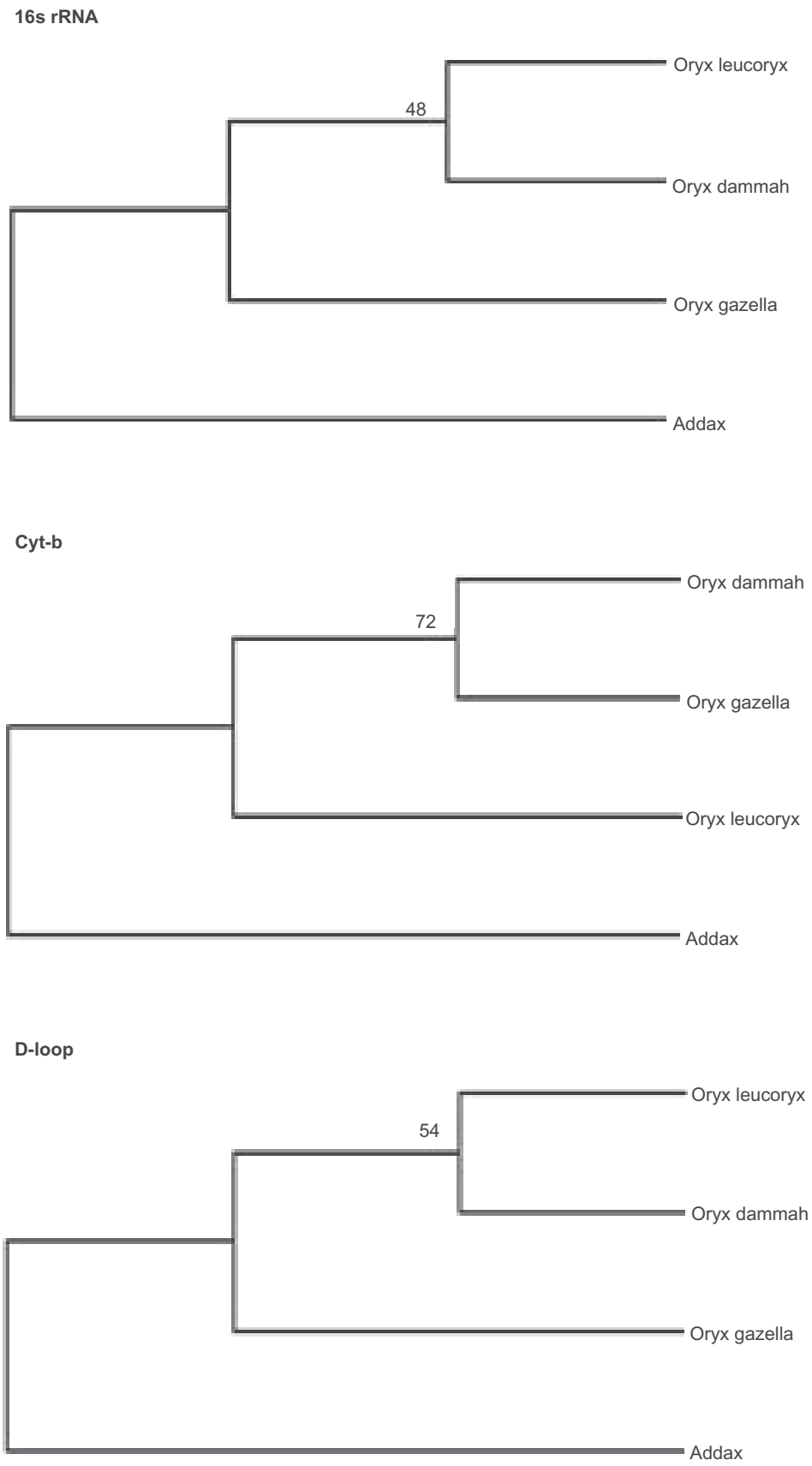

Figure 3. UPGMA method for inferring phylogenetic relationship among various Oryx species using Addax as an outgroup. The bootstrap consensus trees inferred from 1000 replicates are taken to represent the phylogeny. The evolutionary distances were computed using the maximum composite likelihood method.

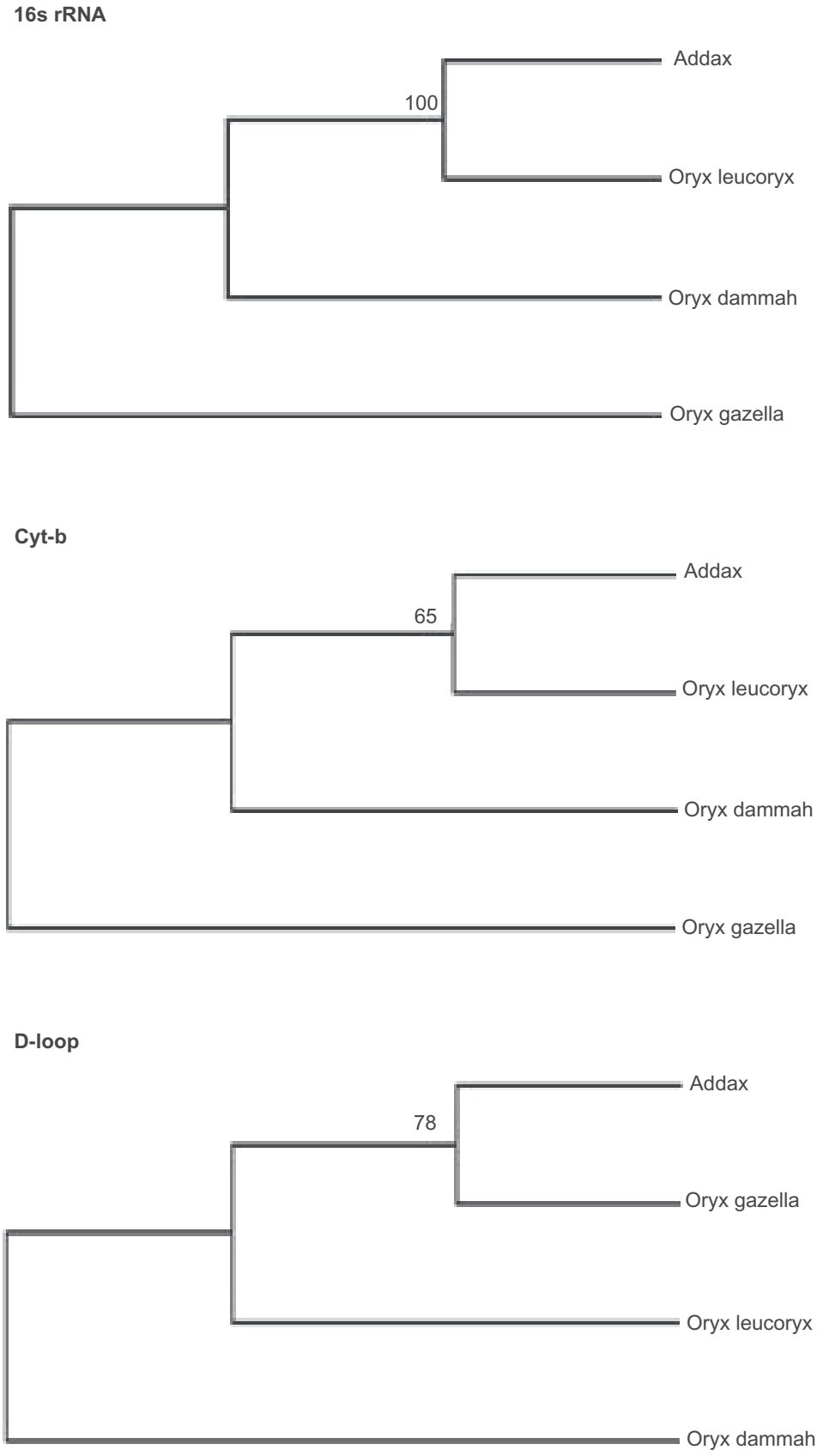

Figure 4. Maximum parsimony method for inferring phylogenetic relationship among various Oryx species using Addax as an outgroup. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed. The maximum parsimonious tree was obtained using the close neighbor interchange algorithm in which the initial trees were obtained with the random addition of sequences (10 replicates).

noticed earlier (Jogger and Garrido, 2001). Shoup and Lewis (2003) have performed BA analyses as well as ML bootstrapping and revealed several instances of conflict between these two approaches to measuring edge support. Kim et al. (2006) have also observed some variation in the topologies of BA and ML-based phylogenic trees to explain the origin and evolution of coronaviruses.

In conclusion, this bioinformatics approach demonstrates the superiority of BA and UPGMA models over MP model for phylogenetic analysis using different regions of mtDNA or other datasets of this size. However, the implication of these findings to different data structures e.g. multiple sequences and more numbers of taxa or outgoups is not clear and needs further investigations.

Acknowledgments

This study was supported by a grant from HRH Prince Sultan Bin Abdulaziz to establish a Research Chair Program in Environment and Wildlife at KSU, Riyadh, Saudi Arabia.

Disclosure

The authors report no conflicts of interest.

References

Bruno, W.J. and Halpern, A.L. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.*, 16:564–6.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F. and Douzery, E.J.P. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.*, 20:248–54.

Eck, R.V. and Dayhoff, M.O. 1966. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Springs, Maryland

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–91.

Felsenstein, J. 2001. The troubled growth of statistical phylogenetics. *Syst. Biol.*, 50:465–7.

Frankham, R., Ballou, J.D. and Briscoe, D.A. 2002. *Introduction to Conservation Genetics*. Cambridge University Press, New York.

Gerber, A.S., Loggins, R., Kumar, S. and Dowling, T.E. 2001. Does non-neutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes. *Ann. Rev. Genet.*, 35:539–66.

Hassanin, A. and Douzery, E.J.P. 1999. The tribal radiation of the family Bovidae (Artiodactyla) and the evolution of the mitochondrial cytochrome b gene. *Mol. Phylogenet. Evol.*, 13:227–43.

Henderson, D.S. 1974. Were they the last Arabian Oryx. *Oryx*, 12:347–50.

Huelsenbeck, J.P. and Ronquist, F.R. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17:754–5.

Ingman, M., Kaessmann, H., Pääbo, S. and Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708–13.

Iyengar, A., Diniz, F.M., Gilbert, T., Woodfine, T., Knowles, J. and Maclean, N. 2006. Structure and evolution of the mitochondrial control region in oryx. *Mol. Phylogenet. Evol.*, 40:305–14.

Iyengar, A., Gilbert, T., Woodfine, T., Knowles, J.M. et al. 2007. Remnants of ancient genetic diversity preserved within captive groups of scimitar-horned oryx (Oryx dammah). *Mol. Ecol.*, 16:2436–49.

Joger, U. and Garrido, G. 2001. Phylogenetic position of Elephas, Loxodonta and Mammuthus, based on molecular evidence. The World of Elephants—International Congress, Rome, 544–7.

Kim, O.J., Lee, D.H. and Lee, C.H. 2006. Close Relationship Between SARS-Coronavirus and Group 2 Coronavirus. *J. Microbiol.*, 44:83–91.

Kumar, S. and Gadagkar, S.R. 2001. Disparity Index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics*, 158:1321–7.

Land, D.E. and Lacy, R.C. 2000. Introgression level achieved through Florida panther genetic restoration. *Endang. Spec. Upd.*, 17:99–103.

Liberles, D.A. 2005. Using phylogeny to understand genomic evolution. In:Parsimony, phylogeny and genomics (ed. Albert, V. A.) pp. 181–19, Oxford University Press, U.K.

Mar, J.C., Harlow, T.J. and Ragan, M.A. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC. Evol. Biol.*, 5:1–20.

Masembe, C., Muwanika, V.B., Nyakaana, S., Arctander, P. and Siegismund, H.R. 2006. Three genetically divergent lineages of the oryx in eastern Africa: Evidence for an ancient introgressive hybridization. *Conserv. Genet.*, 7:551–62.

Mésochina, P., Bedin, E. and Ostrowski, S. 2003. Reintroducing antelopes into arid areas: lessons learnt from the oryx in Saudi Arabia. *C. R. Biol.*, 326:S158–S165.

Montgomery, M.E., Ballou, J.D., Nurthen, R.K., England, P.R., Briscoe, D.A. and Frankham, R. 1997. Minimizing kinship in captive breeding programs. *Zoo. Biol.*, 16:377–89.

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

Olivo, P.D., Van de Walle, M.J., Laipis, P.J. and Hauswirth, W.W. 1983. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature*, 306:400–2.

Ostrowski, S., Bedin, E., Lenain, D. and Abuzinada, A.H. 1998. Ten years of Arabian oryx conservation breeding in Saudi Arabia—achievements and regional perspectives. *Oryx*, 32:209–22.

Page, R.D.M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. *Comp. Appl. Biosci.*, 12:357–8.

Penny, D., Lockhart, P.J., Steel, M.A. and Hendy, M.D. 1994. The role of models in reconstructing evolutionary trees. In:*Models in phylogenetic reconstruction* (eds. Scotland, R.W., Siebert, D.J. and Williams, D.M.) pp. 211–30, Clarendon Press, Oxford.

Ragan, M.A., Murphy, C.A. and Rand, T.G. 2003. Are Ichthyosporea animals or fungi? Bayesian phylogenetic analysis of elongation factor 1α of *Ichthyophonus irregularis*. *Mol. Phylogenet. Evol.*, 29:550–62.

Russello, M.A. and Amato, G. 2007. On the horns of a dilemma: molecular approaches refine ex situ conservation in crisis. *Mol. Ecol.*, 16:2405–6.

Shoup, S. and Lewis, L. 2003. Polyphyletic origin of parallel basal bodies in swimming cells of chlorophycean green algae (Chlorophyta). *J. Phycol.*, 39:789–96.

Sneath, P.H.A. and Sokal, R.R. 1973. *Numerical Taxonomy*. Freeman, San Francisco.

Spalton, J.A., Lawrence, M.W. and Brend, S.A. 1999. Arabian Oryx reintroduction in Oman: successes and setbacks. *Oryx*, 33:168–75.

Sunnucks, P. 2000. Efficient genetic markers for population biology. *TREE*, 15:199–203.

Tajima, F. 1989. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–95.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, 24:1596–9.

Tamura, K., Nei, M. and Kumar, S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U.S.A.*, 101:11030–5.

Westemeier, R.L., Brawn, J.D., Simpson, S.A., Esker, R.W., Jansen, R.W., Walk, J.W., Kershner, E.L., Bouzat, J.L. and Paige, K.N. 1998. Tracking the long-term decline and recovery of an isolated population. *Science*, 282:1695–8.

Xiong, J. and Bauer, C.E. 2002. A cytochrome b origin of photosynthetic reaction centers: an evolutionary link between respiration and photosynthesis. *J. Mol. Biol.*, 322:1025–37.