Accuracy of Estimating Genetic Distance between Species from Short Sequences of Mitochondrial DNA

Andrew P. Martin, Bailey D. Kessing, and Stephen R. Palumbi
Department of Zoology and Kewalo Marine Laboratory, University of Hawaii

Recent advances in molecular techniques allow rapid determination of genetic divergence, and in many circumstances the polymerase chain reaction (PCR) and DNA sequencing have become the technologies of choice for estimating genetic distance. For evolutionary studies, considerable effort is being focused on "universal" oligonucleotide primers that amplify short sequences of mitochondrial DNA. Kocher et al. (1989, p. 6199) encapsulate this approach when they state that a "short sequence from a piece of the cytochrome b gene contains phylogenetic information extending from the infraspecific level to the intergeneric level." If this is a valid generalization, then the advent of universal primers and the ease of direct sequencing of PCR products provides an efficient and reliable means of determining genetic divergence and inferring phylogenetic relationships among taxa. Indeed, DNA sequence derived from the cytochrome b primers (usually amounting to 250 bp) have proved successful for inferring evolutionary relationships and estimating rates and processes of molecular evolution (Kocher et al. 1989; Thomas et al. 1989). However, before evolutionary biologists turn their attention to such short DNA sequences, it is important to understand the accuracy of genetic distances estimated from such sequences.

It is well known that the distribution of substitution events is leptokurtic. Thomas and Beckenbach (1989) showed that there is spatial heterogeneity in the distribution of substitutions in mitochondrial protein-coding genes from salmonid fishes. Nonrandom distribution of substitutions is a problem in estimating evolutionary relatedness from short, amplified DNA sequences because the sampling scheme is nonrandom, being entirely dependent on the choice of primers used for amplifications. Amplification and sequencing of fragments that are too small to encompass the scope of spatial heterogeneity in the distribution of substitutions will give biased estimates of genetic distance.

We investigated the influence of the number of basepairs on genetic distance estimates based on mtDNA sequence data by randomly subsampling known DNA sequences and then determining the variance of genetic distance estimates among subsamples. This approach allows us to investigate the extent to which substitution differences in different parts of a gene interfere with the robustness of genetic distance estimates based on short sequences.

The accuracy of between-taxa genetic divergence estimation based on a small number of basepairs was evaluated by subsampling regions of the mitochondrial genomes containing protein-coding genes for five pairs of vertebrate taxa (fig. 1). The analysis focuses on fourfold-degenerate sites within protein-coding regions because these sites are free from selective constraints and therefore provide the best type of data for inferring evolutionary distance (Wu and Li 1985).

For each of the five paired vertebrate comparisons, continuous, homologous sequences of 250, 500, 750, 1,000, 1,250, and 1,500 bp were selected at random from the complete sequences. For these sequence lengths, the average numbers of fourfold-

Address for correspondence and reprints: Andrew P. Martin, Department of Zoology and Kewalo Marine Laboratory, University of Hawaii, Honolulu, Hawaii 96822

Mol. Biol. Evol. 7(5):485–488. 1990.

0737-4038/90/0705-0009$02.00

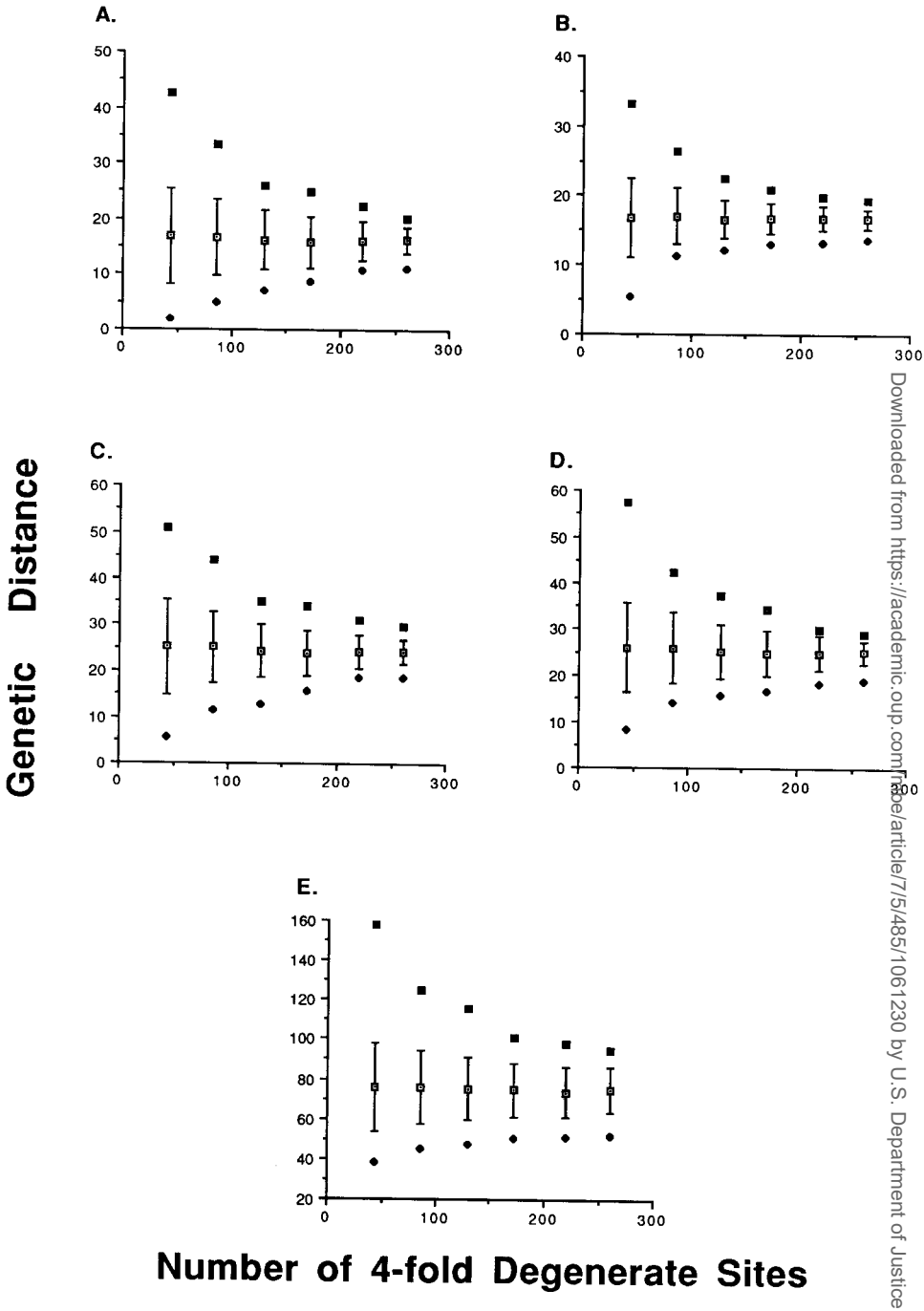

FIG. 1.—Mean (⊡), 1 SD (vertical lines), and minimum (♦) and maximum (■) estimates of genetic distance from subsampling, for the five pairs of taxa compared. Genetic distance is expressed as K4 × 100 and is therefore equivalent to percent sequence divergence. A, Rainbow-cutthroat trout (*Salmo gardineri–S. clarkii;* average divergence 16.2%). B, Coho-chinook salmon (*Oncorhyncus kisutch–O. tschawytscha;* average divergence 16.8%). C, Cutthroat-sockeye salmon (*S. clarkii–O. nerka;* average divergence 24.1%). D, Rainbow-chinook salmon (*S. gardineri–O. tschawytscha;* average divergence 25.2%). E, Rat-mouse (*Rattus norvegicus–Mus musculus;* average divergence 74.4%). The salmonid data consist of 2,067 bp from the ATPase 6, CO III, ND 3, and ND 4L genes (Thomas and Beckenbach 1989). The rat and mouse data comprise 5,085 bp from the CO I, CO III, cytochrome b, ND 1, and ATPase 6 genes (Bibb et al. 1981; Gadaleta et al. 1989).

degenerate sites sampled were 43, 85, 130, 172, 220, and 261, respectively. For each paired comparison, the number of base substitutions at fourfold-degenerate sites was determined using $K4 = 0.5 \times \ln[1/(1-2P-Q)] - 0.25 \times \ln[1/(1-2Q)]$, where P and Q are the proportions of transitional and transversional differences, respectively (Wu and Li 1985). This process was repeated 400 times for each of the subsample sizes, and the means and variances of the K4 values were determined. A limitation of this approach is that repetitive subsampling of the same larger sample results in subsamples that are not completely independent. This reduces the variance, and so our variance estimates should be considered as minima.

In addition, for each subsample size, the variance in the estimate of K4 was calculated for 100 replicates by using the approximation equation:

$$\mathrm{Var}(K4) = [a^2P + d^2Q - (aP+dQ)^2]/L, \quad (1)$$

where $a = 1/(1-2P-Q)$, $b = 1/(1-2Q)$, $d = (a+b)/2$, and L is the number of fourfold-degenerate sites (Wu and Li 1985). For each subsample size, the average of the 100 standard deviations (SDs) determined using equation (1) was compared with the SD of the K4 estimates from subsampling.

As expected, mean estimates of divergence at synonymous sites do not vary greatly with the number of bases sampled (fig. 1). The mean number of synonymous substitutions of 400 replicate samples is not statistically different from the value for synonymous substitutions for the whole sequence. However, the range and SD of synonymous substitutions among subsample replicates show pronounced decline with increasing sample size (fig. 1).

In the majority of cases, the SD of K4 values from subsampling exceeded, by 30%–40%, the SD predicted from equation (1). These results suggest that variances in number of substitutions estimated from theory are underestimates of the true variance in the pattern of substitutions present in DNA sequences.

Regression analysis of the coefficient of variation (CV) against subsample size provides an indication of the sampling effort required to obtain a desired level of accuracy. Logarithmic regression equations are fitted to the two extreme sets of points, providing boundaries to evaluate the dependence of CV on the number of basepairs sampled. The results show (fig. 2) that, for sequences of ~250 bp or less, the CV is usually >50. For sequences of 250–500 bp, CV tends to be 30–50. For sequences >750 bp in length, CV can be as low as 10 or as high as 30.

Figure 2 also shows that there are important differences between data sets in the CVs resulting from subsampling. For example, the rainbow-cutthroat comparison gives a much higher CV than does the coho-chinook comparison, despite similar levels of divergence (average divergence 16.2% and 16.8%, respectively). The primary difference between these two comparisons is that the spatial variation of substitutions is more extreme for the rainbow-cutthroat comparison than for the coho-chinook comparison: variances for the number of codons separating variable fourfold-degenerate nucleotide positions are 165.5 and 92.6, respectively.

There are two possible solutions to the problem of short DNA sequences. Because gross errors in estimation of evolutionary relatedness can occur when calculations of genetic distance are based on few basepairs, the most straightforward approach is to sequence larger sections of DNA. The actual level of sampling effort required is dependent on the underlying distribution of substitutions. Marked departures from a random spatial distribution of substitutions demand greater sampling efforts. The second method is to sequence a small piece of mtDNA from numerous individuals or species that have diverged at the same time. Mutational changes along lineages in this "star phylogeny" (Gillespie 1984) will be independent, and so data may be taken together to yield a good estimate of *average* genetic distance among all lineages used.

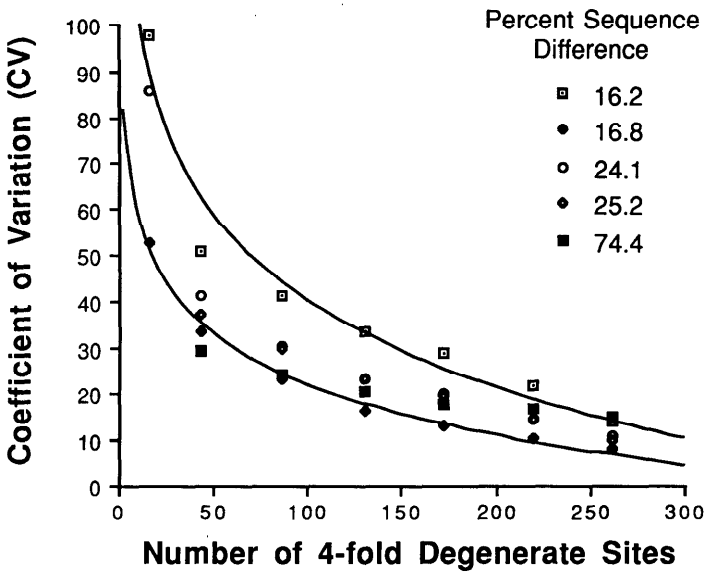

FIG. 2.—Graph of CV, plotted against the number of fourfold-degenerate basepairs sampled for the five paired comparisons. Logarithmic regression equations are fitted to the lowest and highest CV values.

Acknowledgments

This research was supported by NSF grant BSR 8604969 and by a Whitehall Foundation grant, both to S.R.P.

LITERATURE CITED

BIBB, M. J., R. A. VAN ETTEN, C. T. WRIGHT, M. W. WALBERG, and D. A. CLAYTON. 1981. Sequence and gene organization of the mouse mitochondrial DNA. Cell **26**:167–180.

GADALETA, G., G. PEPE, G. DE CANDIA, C. QUAGLIARIELLO, E. SBISA, and C. SACCONE. 1989. The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. J. Mol. Evol. **28**:497–516.

GILLESPIE, J. H. 1984. The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. USA **81**:8009–8013.

KOCHER, T. D., W. K. THOMAS, A. MEYER, S. V. EDWARDS, S. PAABO, F. X. VILLABLANCA, and A. C. WILSON. 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA **86**:6196–6200.

THOMAS, R. H., W. SCHAFFNER, A. C. WILSON, and S. PAABO. 1989. DNA phylogeny of the extinct marsupial wolf. Nature **340**:465–467.

THOMAS, W. K., and A. T. BECKENBACH. 1989. Variation in salmonid mitochondrial DNA: evolutionary constraints and mechanisms of substitution. J. Mol. Evol. **29**:233–245.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

WALTER M. FITCH, reviewing editor

Received January 5, 1990; revision received May 8, 1990

Accepted May 8, 1990