Submitted to
Mol. Biol. and Evolution, 1985.
(Now under revision).

On estimating substitution rates from pairs of homologous nucleotide sequences

Simon Tavaré* and Tamra Janzen

Department of Statistics

Colorado State University

Fort Collins

CO 80523

*to whom proofs should be sent.

Running Head: Estimating Substitution Rates

Keywords: *Substitutions. Divergence. Mitochondrial genome.*

ABSTRACT

This paper describes some methods for estimating the mean number of substitutions that have occured in two homologous nucleotide sequences since their divergence from a common ancestor. The novel ingredients allow for arbitrary ancestral composition and possibly different substitution matrices for the two species. The methods are illustrated by mitochondrial nucleotide sequences from mouse and man. Evidence of different substitution rates in the two species is found, although the estimates of the total number of substitutions are similar to those found by previous methods.

INTRODUCTION

Several authors have considered the problem of estimating the mean number of substitutions per site since divergence of a pair of functionally homologous nucleotide sequences. For recent approaches see Tajima and Nei (1984), Lanave et al. (1984), and the reviews of Kaplan (1983) and Tavare (1985). The basic data used in such studies are as follows. Consider two functionally homologous nucleotide sequences of length n taken one from each of two species. The bases in the sequences are labelled 1, 2, 3, 4 for A, T, C, G respectively. On the basis of the aligned sequences we calculate the numbers N_{ij} defined by

N_{ij} = number of times an aligned site has a base
of type i in species 1 and a base of type j (1)
in species 2.

This results in a 4 x 4 matrix $N = (N_{ij})$ of observations.

We assume that the species in question diverged from a common ancestor T years ago and behave independently after divergence. The bases in each sequence are changed through time by the effects of substitutions. Under these assumptions, the probability f_{ij} that a site has a base of type i in species 1 and of type j in species 2 is given by

$$f_{ij} = \sum_{\ell=1}^{4} \pi_\ell \, p^1_{\ell i} \, p^2_{\ell j} \quad , \tag{2}$$

where π_ℓ is the probability that the ancestral base is of type ℓ, and $p^r_{\ell i}$ is the probability that in species r (r = 1 or 2) a base of type ℓ at divergence is of type i a time T later. Most authors have used (implicitly or explicitly) a Markov chain model to specify the probabilities $p^r_{\ell i}$. That is we specify matrices $Q_1 = (q^1_{ij})$ and $Q_2 = (q^2_{ij})$ that satisfy

$$q^r_{ij} \geq 0 \ (i \neq j) \ : \ q^r_{ii} = - \sum_{j \neq i} q^r_{ij} \ , \ r = 1,2 \qquad (3)$$

and then

$$P_r := (p^r_{ij}) = \exp(Q_r T) \ , \ r = 1,2 \qquad (4)$$

where

$$\exp(Q_r T) := \sum_{n=0}^{\infty} Q^n_r \frac{T^n}{n!} \ .$$

(cf. Karlin and Taylor (1975), p. 150) The assumptions made by other authors include:

A1. $Q_1 = Q_2 =: Q$, implying that $P_1 = P_2$

A2. $\underline{\Pi}' \ Q = \underline{0}$, implying that $\underline{\Pi} = (\Pi_1, \Pi_2, \Pi_3, \Pi_4)$ is the stationary distribution of Q, and then that the distribution of base composition in each species does not vary with time.

Notice that assumption A1 means that the matrix $F = (f_{ij})$ defined in (2) is symmetric; the data matrix N should reflect such symmetry. There is ampple evidence that this is not the case, particularly for data that arise from the third codon positions; cf. Tavare (1985). In this paper we will not assume that the rate matrices in the two species are equal, or that the ancestral frequencies $\underline{\Pi}$ are stationary. We will describe a method for estimating parameters of such asymmetric models, and use it to estimate the mean number of substitutions that have occurred (per base) in each species since divergence. We will also investigate whether the mean number of substitutions in each species are equal. As a final application, we calculate the probability that a site has not changed, conditional on its having identical nucleotides.

THEORY

The data matrix N is 15 dimensional, while a general Q-matrix satisfying (3) is 12-dimensional. Allowing for three probabilities for the ancestral composition, this gives us 12 + 12 + 3 = 27 parameters to estimate for the general model described by (2) and (4). It follows that we must restrict our class of possible models in some way. We chose to analyse two cases. In the first, which we denote as model (K), each Q-matrix is of the form

$$\text{Model (K)} \qquad Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \left[\begin{array}{cccc} \cdot & \alpha_1 & \alpha & \alpha \\ \beta_1 & \cdot & \alpha & \alpha \\ \beta & \beta & \cdot & \alpha_2 \\ \beta & \beta & \beta_2 & \cdot \end{array}\right] \end{array}$$

This particular form was proposed by Kimura (1981). In the second model, denoted here as model (TK), the Q-matrices are of the form

$$\text{Model (TK)} \quad Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \left[\begin{array}{cccc} \cdot & \gamma & \theta\alpha & \alpha \\ \gamma & \cdot & \alpha & \theta\alpha \\ \theta\beta & \beta & \cdot & \gamma \\ \beta & \theta\beta & \gamma & \cdot \end{array}\right] \end{array} .$$

This example was suggested by Takahata and Kimura (1981). In both cases, the diagonal elements are found from the requirement that the rows sums be zero; see (3).

The statistical problem can now be described as follows. We assume that sites evolve independently of one another, so that the data matrix N has the form of a multinomial trials experiment, in which there are 16 cells, with underlying cell probabilities $F = (f_{ij})$ given by (2). These cell probabilities are functions of p unknown variables, $\underline{x} = (x_1,\ldots,x_p)$, say, which are to be estimated from the data. For model (K), p = 15, while for model (TK), p = 11. We can estimate these parameters by maximum likelihood or minimum chi-squared methods. Numerically, we obtained similar results in both cases; only the maximum likelihood results are described here. There are some difficult computational aspects associated with this problem, since closed-form solutions of the likelihood equations seem impossible to find. Details of the computational approach are given in the Appendix. We also computed the approximate variance-covariance matrix of the estimates $\hat{\underline{x}}$ of $\underline{x}$ for large sequence length n. This matrix is then used to compute variance estimates for quantities of interest later on. See appendix. The theory of maximum likelihood estimation in multinomial trials is well documented; a particularly accessible account is given in Cox (1984) for example.

Primarily we are interested in estimating the mean number K_r of substitutions that have occurred in species r since divergence from the common ancestor. It can be shown that for models of the type considered here, we get

$$K_r = \sum_i \Pi_i \sum_j q_j^r \int_0^T \left[e^{Q_r s}\right]_{ij} ds, \; r = 1,2 . \qquad (5)$$

In equation (5), $q_j^r := -q_{jj}^r$. If indeed $\underline{\Pi}'Q_r = \underline{0}'$, so that $\underline{\Pi}$ is the equilibrium distribution for species r, then (5) reduces to the more familiar

$$K_r^e = T \sum_i \pi_i \, q_i^r, \tag{6}$$

'e' denoting 'equilibrium'. It is this latter parameter that other authors have estimated; cf. equation (19) of Lanave et al. (1984), or equation (1) of Tajima and Nei (1984). It is important to note that the divergence time T itself is confounded with other parameters in (5) and (6), and so cannot be separately estimated without additional information. In this paper, we have adopted the convention that T = 1, so the equation (6) becomes $K_r^e = \Sigma \pi_i \, q_i^r \; (\equiv \Sigma \pi_i (q_i^r \; T))$, for example. The estimated Q-matrices given in Tables 1,2 and 7 are then estimates of Q_1T and Q_2T and so on.

It appears to be very hard to prove that a unique solution to the likelihood equations exists. Further, the numerical analysis routine often found elements of the Q-matrices that were algorithmically zero. We therefore adopted the following approach. We started the optimisation algorithm from a variety of different starting points (for example, using different initial estimates of the ancestral frequencies), and compared the results. Any parameters in the Q-matrices that were computationally zero (say, less than 10^{-6}) were set to zero, and not used as parameters. This has the effect of reducing the dimension p of the problem, and thus increases the degrees of freedom for the goodness-of-fit of the model to the data.

AN EXAMPLE

The data used for this example are taken from the EMBL sequence library. The sequences are from mouse (species 1) and human (species 2) mitochondrial genomes; the former is from Bibb et al. (1981), the latter from Anderson et al. (1982). As done by Lanave et al. (1984), we considered the nucleotide sequence of the five mt genes coding for identified products, namely the three cytochrome oxidase subunits (COI, COII, COIII), the ATPase subunit and the cytochrome b (cyt b) subunit. The total length of the resulting "super sequence" is 4803 nucleotides. The data discussed below comprises only the third codon position sequence, resulting in a length of n = 1601 nucleotides. The data matrix N is

$$N = \begin{array}{ccccc} & A & T & C & G \\ A & 386 & 74 & 193 & 65 \\ T & 83 & 104 & 216 & 6 \\ C & 76 & 81 & 265 & 9 \\ G & 24 & 3 & 12 & 4 \end{array}$$

It is clear that the data matrix is not consistent with symmetry (as would be required if assumption A1 were valid). A formal Chi-squared test of symmetry (cf. Bowker (1948)) had a value of χ^2 = 133, with six degrees of freedom (d.f.) We then fitted models (K) and (TK) to the data. The estimated Q-matrices are given in Tables 1 and 2.

INSERT TABLE 1 HERE

INSERT TABLE 2 HERE

Qualitatively, these results are similar. The chi-squared statistics for goodness-of-fit were 20.78 (model(K); 6 d.f.) and 12.94 (model (TK); 6 d.f.); model (TK) seems to be a better description of the data. One suprising point of our analysis involves the estimated ancestral frequencies, $\underline{\Pi}$; these are given in Table 3.

INSERT TABLE 3 HERE

The average nucleotide composition of the present day sequences is .402 (A), .210(T,, .348(C) and .040(G). The estimated ancestral frequencies suggest that nucleotide T started from very low frequency, and that the present observed frequency is due to non-stationarity in the substitution process, in contrast to the 'usual' assumption.

Table 4 records the observed present-day nucleotide frequencies obtained from the data matrix N, and the corresponding frequencies calculated from (2) for each of the models.

INSERT TABLE 4 HERE

From the results it is clear that the two marginal distributions are not equal (this is really a reflection of the asymmetry in the data, but can assessed statistically using a test of marginal homogeneity; cf. Tavaré (1985)). The predicted frequencies are in good agreement with those observed.

Next, we estimate the mean number of substitutions per base in each species since divergence. This is calculated from (5); the results are given in Table 5.

INSERT TABLE 5 HERE

The estimated standard error of the difference between $\hat{K}_1$ and $\hat{K}_2$ is .036 (for model (K)) and .040 (for model (TK)). Both of these results suggest that there is a significant difference between the substitution processes in each species; the rates of substitution are different.

It is also worth comparing the estimates $\hat{K} = \hat{K}_1 + \hat{K}_2$ of the average number of substitutions per site since divergence with those based on previous models in which assumptions A1 and A2 are used. Some representative results are given in Table 6. Despite the different type of assumptions used in these models, the results are remarkably consistent.

INSERT TABLE 6 HERE

The data used here were based on almalgamating five coding regions from the mitochondrial genomes. To assess possible inhomogeneity in the sequence, we analysed the third base sequence (of 512 nucleotides) of the Cytochrome Oxidase I subunit using the models described above. In Tables 7 and 8 the Q-matrices and ancestral frequencies for models (K) and (TK) are given. Qualitatively, these results are the same as for the combined sequence; further comparisons of the Q-matrices appear in the next section.

INSERT TABLE 7 & 8 HERE

For COI, the estimates of mean substitution number analogous to those of Table 5 are given in Table 9.

INSERT TABLE 9 HERE

The estimated standard error of $\hat{K}_1 - \hat{K}_2$ is .08 for model (TK) and .07 for model (K); both these results are consistent with different substitution rates. The estimates of $\hat{K} = \hat{K}_1 + \hat{K}_2$ for the number of substitutions per site are $\hat{K} = .88 \pm .08$ for model (K), and $\hat{K} = 1.03 \pm .08$ for model (TK).

THE NUMBER OF UNCHANGED SITES

In this section we will estimate for sites which currently are identical the proportion which have never changed. That is, we estimate the probabilities η_i, i = 1,2,3,4 defined by

η_i = Pr (a site has never been substituted, given that the site now has both nucleotides of type i).

Under our model, this may be calculated as

$$\eta_i = \Pi_i \exp\{-(q_i^1 + q_i^2)T\} / f_{ii} \tag{7}$$

In Table 10, we give the values of $\hat{\eta}_i$ obtained by substituting estimates of the corresponding parameters into (7). It will be observed that these probabilities vary widely among nucleotides, although they are fairly consistent between models, and between the COI subunit and the full data set. As a summary of this data, we also computed the probability η that, given a site is identical, no substitutions have occurred at it. This can be derived analogously to (7) to give

$$\eta = \frac{\sum_i \Pi_i \exp\{-(q_i^1 + q_i^2)T\}}{\sum_i f_{ii}} . \tag{8}$$

The estimated values of η are given in Table 11.

INSERT TABLE 10 HERE

INSERT TABLE 11 HERE

The results indicate that about 80% of all sites showing identical nucleotides have never received a substitution since divergence. Of course, this should not be taken as a measure of the fraction of sites that can never change, since under this model all sites must eventually receive at least one substitution. It does provide one way of comparing the different Q-matrix estimates, though.

SUMMARY

The methods presented here provide a class of models for estimating the mean number of substitutions per site that have occurred since divergence from a common ancestor. These methods are computational (rather than analytic) in style. They allow for arbitrary ancestral frequencies for the nucleotides; these may be estimated from the data (as was the case in the results presented here) or may be taken as fixed and known. We have not recorded the details of results for this latter case. Suffice it to say that if the ancestral composition is taken to be the average present-day composition (as is often the case in other studies of this problem) then no satisfactory fit to the data can be obtained for either model (K) or model (TK). For the particular data set used here, there was evidence that the substitution rates in the mouse and human mitochondrial genomes are different, although the estimated mean number of substitutions per site agrees substantially with the results of previous methods. There was a significant non-homogeneity in the composition of the nucleotide frequencies over time; the estimated ancestral frequencies indicate that the G and T nucleotides have increased substantially in frequency since divergence. The data used here combined the five coding regions into one large sequence. In order to check the consistency of the results, we ran analogous algorithms on the COI subunit. The results were qualitatively similar to those obtained for the whole sequence. As a way to compare the estimated substitution schemes, we computed the probability that a

site now showing identical bases has never had a substitution; this probability was estimated at about 80%.

The methods developed here were restricted to a small subclass of a set of possible models; this restriction was required by the number of data cells which could be used to estimate parameters. If the data consisted of (say) three species, then more general models can be fitted since restriction on parameter numbers vanish. What is important, though, is that precisely the same techniques apply to that problem as have been developed here.

APPENDIX

COMPUTATIONAL DETAILS

The maximum likelihood estimators of the parameter $\underline{x}$ of the model are obtained by minimising the function $L(\underline{x})$ given by

$$L(\underline{x}) = - \sum_{i,j} N_{ij} \log f_{ij}(\underline{x})$$

where the f_{ij} are defined by equations (2), (3) and (4). These functions must be minimised with several constraints operating; for example, the off-diagonal elements of the two Q-matrices must be positive (≥ 0), and the estimated ancestral frequencies $\underline{\Pi}$ must satisfy $\Pi_i \geq 0$, $i = 1,2,3,4$ and $\Pi_1 + \Pi_2 + \Pi_3 + \Pi_4 = 1$. Our approach was to reparameterise the problem so as to produce a new model in which the parameter were unconstrained (for example, if the constraint $x \geq 0$ is required, set $x = e^y$ or $x = y^2$; y is now unconstrained.) The new unconstrained problem was approached using the IMSL subroutine library program ZXMIN, and the answerrs rescaled to give a solution $\hat{\underline{x}}$ of the original problem. We are unable to establish analytically whether a unique solution of the likelihood equations exists. We started the optimisation algorithm from a variety of initial positions (typically, 5 due to the computer-time required). In all cases reported here, the same final solution was obtained. For the data examined here, several parameter values were computationally zero, and so occurred on the boundary of the parameter space. Our approach was to set any such parameter (say, with a value of less than 10^{-6}) to zero, effectively reducing the number of parameters to be estimated by one.

Part of the algorithmic difficulty in this problem is the evaluation of matrix exponentials of the type required by (2) and (4); cf. Moler and Van Loan (1978). For the model (K), an explicit formula for exp(QT) is available from Gojobori et al. (1982); we used their method. For model (TK), we computed the matrix exponential by a diagonalisation method. If this failed, we resorted to an efficient series mehtod which converged rapidly.

The asymptotic standard deviations quoted here relied on the computation of the variance-covariance matrix $\frac{1}{n}\Sigma$ of the estimated parameters $\hat{\underline{x}}$. The r-sth element of Σ^{-1} is given by

$$C_{rs} = \sum_{i,j} \frac{1}{f_{ij}(\underline{\hat{x}})} \frac{df_{ij}(\underline{\hat{x}})}{dx_s} \frac{df_{ij}(\underline{\hat{x}})}{dx_r}$$

We evaluated all derivatives by high-order forward difference formulae, using the same step-size algorithm as IMSL's ZXMIN. The joint asymptotic distribution of $(\hat{K}_1, \hat{K}_2) = (\hat{K}_1(\underline{\hat{x}}), \hat{K}_2(\underline{\hat{x}}))$ is multivariate normal, with variance-covariance matrix V given by

$$V = \frac{1}{n} D\Sigma D^T$$

where $D_{ir} = \dfrac{dK_i(\underline{\hat{x}})}{dx_r}$, $i = 1,2;\ r = 1,\ldots,p.$

Once more, the derivatives required above were computed numerically. We would be happy to provide further details of the numerical methods to anyone who is interested.

LITERATURE CITED

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. L. DeBruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, and I. G. Young. 1981. Sequence and organisation of the human mitochondrial genome. Nature, 290, 457-465.

Bibb, M. J., R. A. Van Etten, C. T. Wright, M. W. Walberg and D. A. Clayton, 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell. 26, 167-180.

Bowker, A. H., 1948. A test for symmetry in contingency tables. J. Amer. Statist. Soc., 43: 572-574.

Cox, C., 1984. An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. American Statistician, 38, 283-287.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17: 368-376.

Gojobori, T., K. Ishii and M. Nei, 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol., 18, 414-423.

Jukes, T. H. and Cantor, C. H., 1969. Evolution of protein molecules in mammalian protein metabolism. H. N. Munro Ed., Academic Press, New York, pp. 21-123.

Kaplan, N., 1983. Statistical analysis of restriction enzyme map data and nucleotide sequence data. In statistical analysis of DNA sequence data, B. S. Weir Ed., Dekker, New York, pp. 75-107

Kaplan, N. and K. Risko, 1982. A method for estimating rates of nucleotide substitution using DNA sequence data. Theor. Popn. Biol., 21: 318-328.

Karlin, S. and H. M. Taylor, 1975. A first course in stochastic processes. 2nd Edn. Academic Press, New York.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. 78, 454-458.

Lanave, C., G. Preparata, C. Saccone and G. Serio, 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol., 20, 86-93.

Moler, C. and C. Van Loan, 1978. Nineteen dubious ways to compute the exponential of a matrix. SIAM Review, 20, 801-836.

Tajima, F. and M Nei, 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol., 1, 269-285.

Takahata, N. and M. Kimura, 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics, 98, 641-657.

Tavaré, S., 1985. Some probabilistic and statistical problems in the analysis of DNA sequences. In R. M. Miura, ed. Lectures on mathematics in the life sciences, Vol. 17. Amer. Math. Soc., in press.

TABLE 1

Estimated Q-matrices for model (K)

Species 1 (Mouse)

$$Q_1 = \begin{array}{cc} & \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \left[\begin{array}{cccc} -.205 & .076 & .065 & .065 \\ .000 & -.129 & .065 & .065 \\ .454 & .454 & -.908 & .000 \\ .454 & .454 & .000 & -.908 \end{array}\right] \end{array}$$

Species 2 (Human)

$$Q_2 = \begin{array}{cc} & \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \left[\begin{array}{cccc} -.320 & .056 & .132 & .132 \\ .000 & -.265 & .132 & .132 \\ .232 & .232 & -.464 & .000 \\ .232 & .232 & .000 & -.464 \end{array}\right] \end{array}$$

TABLE 2

Estimated Q-matrices for model (TK)

Species 1
(Mouse)

$$Q_1 = \begin{array}{cc} & \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \begin{bmatrix} -.135 & .000 & .055 & .080 \\ .000 & -.135 & .080 & .055 \\ .407 & .592 & -.999 & .000 \\ .592 & .407 & .000 & -.999 \end{bmatrix} \end{array}$$

Species 2
(Human)

$$Q_2 = \begin{array}{cc} & \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \begin{bmatrix} -.285 & .000 & .122 & .163 \\ .000 & -.285 & .163 & .122 \\ .237 & .319 & -.556 & .000 \\ .319 & .237 & .000 & -.556 \end{bmatrix} \end{array}$$

TABLE 3

Estimated ancestral frequencies $\underline{\Pi}$

Model	Base:	A	T	C	G
(K)		.353	.039	.607	.001
(TK)		.317	.013	.670	.000

TABLE 4

Observed present-day nucleotide frequencies, and their estimates from models (K) and (TK)

	Mouse			
	A	T	C	G
Observed	.448	.256	.269	.027
Model (K)	.460	.244	.270	.026
Model (TK)	.448	.255	.270	.027
	Human			
	A	T	C	G
Observed	.355	.164	.428	.053
Model (K)	.362	.157	.431	.050
Model (TK)	.355	.163	.429	.053

TABLE 5

Estimated mean number of substitutions per base in each species since divergence. Figure after ± is one standard deviation

	Mouse	Human
Model (K)	$\hat{K}_1 = .49 \pm .03$	$\hat{K}_2 = .39 \pm .03$
Model (TK)	$\hat{K}_1 = .52 \pm .03$	$\hat{K}_2 = .44 \pm .03$

TABLE 6

Estimated mean number of substitutions $\hat{K}$ per site since divergence. Figure after $\pm$ is one standard deviation

Method	$\hat{K} \pm$ std. dev.
Model (K)	.88 ± .04
Model (TK)	.96 ± .04
JC(†)	.91 ± .04
KR(††)	.99 ± .05
F(†††)	1.03 ± .06

† Jukes and Cantor (1969)

†† Kaplan and Risko (1982)

††† Felsenstein (1981); Tavare (1985)

TABLE 7

Estimated Q-matrices for model (K) for COI subunit

Species 1
(Mouse)

$$Q_1 = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & -.298 & .112 & .093 & .093 \\ T & .000 & -.186 & .093 & .093 \\ C & .481 & .481 & -.962 & .000 \\ G & .481 & .481 & .000 & -.962 \end{array}$$

Species 2
(Human)

$$Q_2 = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & -.344 & .118 & .113 & .113 \\ T & .000 & -.226 & .113 & .113 \\ C & .183 & .183 & -.366 & .000 \\ G & .183 & .183 & .000 & -.366 \end{array}$$

	A	T	C	G
Ancestral Frequencies	.404	.051	.545	.000

TABLE 8

Estimated Q-matrices for model (TK) for COI subunit

Species 1 (Mouse)

$$Q_1 = \begin{array}{cc} & \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \left[\begin{array}{cccc} -.200 & .000 & .076 & .124 \\ .000 & -.200 & .124 & .076 \\ .432 & .705 & -1.137 & .000 \\ .705 & .432 & .000 & -1.137 \end{array}\right] \end{array}$$

Species 2 (Human)

$$Q_2 = \begin{array}{cc} & \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{array}{c} A \\ T \\ C \\ G \end{array} & \left[\begin{array}{cccc} -.262 & .000 & .102 & .160 \\ .000 & -.262 & .160 & .102 \\ .215 & .336 & -.551 & .000 \\ .336 & .215 & .000 & -.551 \end{array}\right] \end{array}$$

	A	T	C	G
Ancestral Frequencies	.328	.010	.662	.000

TABLE 9

Estimated mean number of substitutions per base in each species since divergence in COI. Figure after ± is one standard deviation

	Mouse	Human
Model (K)	$\hat{K}_1 = .54 \pm .05$	$\hat{K}_2 = .34 \pm .05$
Model (TK)	$\hat{K}_1 = .61 \pm .06$	$\hat{K}_2 = .42 \pm .05$

TABLE 10

Estimated probabilities $\hat{\eta}_i$ that given a site has both nucleotides of type i, that site has never been substituted

		Sequence			
		Whole Data set		COI	
i	Base	Model (K)	Model (TK)	Model(K)	Model (TK)
1	A	.85	.85	.87	.84
2	T	.44	.15	.52	.09
3	C	.93	.91	.91	.88
4	G	.10	.04	.00	.00

TABLE 11

Estimated probabilities $\hat{\eta}$ that, given a site has identical nucleotides, no substitutions have occurred at the site

	Sequence			
	Whole Data Set		CO I	
	Model (K)	Model (TK)	Model (K)	Model (TK)
$\hat{\eta}$	.82	.77	.83	.74