THIRD ROCKY MOUNTAIN
CONFERENCE ON MEDICAL
APPLICATIONS of STATISTICS.
Preprint volume 89-96, 1985.

THE ESTIMATION OF SUBSTITUTION RATES AND DIVERGENCE TIMES FROM DNA SEQUENCE DATA

Simon Tavaré
Statistics Department
Colorado State University
Fort Collins, CO 80523

ABSTRACT

Some models for the estimation of substitution rates from pairs of functionally homologous DNA sequences are compared. A novel feature here, motivated by the observed asymmetry in the data, is that the substituion processes in each arm of the tree are allowed to differ. The statistical method involves maximum likelihood estimation for multinomial trials, whose underlying cell probabilities are determined by continuous-time Markov chains.

I. INTRODUCTION

Consider two functionally homologous nucleotide sequences of length n, taken one from each of two species. The sequences are aligned, to give data of the form:

Species X: A T T C . . . . . G A A
Species Y: G T T C . . . . . G A T.

We form a contingency table, with entries $\{N_{ij}\}$ given by

N_{ij} = number of times an aligned site has a base of type i in species X, and of type j in species Y;

the bases A, T, C, G are labelled 1, 2, 3, 4 respectively. Now assume that the two species in question diverged from a common ancestor T years ago, and that after divergence the two species behaved independently, the bases in each

sequence being changed through time by the substitution of one base for another.

Under these assumptions, the probability f_{ij} that a site has a base of type i in species X, and j in species Y is

$$f_{ij} = \sum_{\ell} s_{\ell} \, p^{X}_{\ell i} \, p^{Y}_{\ell j}, \tag{1.1}$$

where s_{ℓ} is the probability that the ancestral base is ℓ, and $p^{X}_{\ell i}$ (resp., $p^{Y}_{\ell i}$) is the probability that in species X (resp., Y), a base ℓ at divergence is of type i a time T later. Most authors have used a Markov model to specify the probabilities $\{p^{X}_{\ell i}\}$, so that

$$P_X := \{p^{X}_{\ell i}\} = \exp\{Q_X T\}, \tag{1.2}$$

where Q_X is the generator of the (irreducible) X-process. In addition, it is assumed that

(a) $Q_X = Q_Y =: Q$, implying that $P_X = P_Y$

(b) $\underline{s}'Q = \underline{0}'$, implying that $\underline{s} = (s_1, s_2, s_3, s_4)'$ is the stationary distribution of Q. (1.3)

See Lanave et al. (1984), Tajima and Nei (1984) for some recent work in this area. The assumption (1.3a) above means that the cell probability matrix $F = \{f_{ij}\}$ is symmetric; the data matrix N should be consistent with such symmetry. There is ample evidence that this is not the case for many pairs of sequences, particularly for those arising from the third codon positions; Tavaré (1985). This paper therefore focuses on methods of estimation for models like (1.2) in which

(a) Q_X and Q_Y may be different. (1.4)

(b) $\underline{s}$ is not necessarily the stationary distribution of either Q_X or Q_Y.

II. MODELS AND STATISTICAL METHODS

The cell probability matrix F is 15-dimensional, and a general Q-matrix in (1.2) is 12-dimensional. Thus the most general model will have dimension 12 + 12 + 3 = 27.

Thus we need to restrict the form of the Q-matrices to be used.
There are several candidates that might be useful; this paper focusses on just two possibilities.

Model (K) (Kimura (1981), Gojobori et al. (1982))

Here, the Q-matrices are of the form

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & \alpha_1 & \alpha & \alpha \\ \beta_1 & \cdot & \alpha & \alpha \\ \beta & \beta & \cdot & \alpha_2 \\ \beta & \beta & \beta_2 & \cdot \end{pmatrix} \end{array},$$

where the diagonal elements are determined by $Q\underline{1} = \underline{0}$.

Model (TK) (Takahata and Kimura (1981))

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & \gamma & \theta\alpha & \alpha \\ \gamma & \cdot & \alpha & \theta\alpha \\ \theta\beta & \beta & \cdot & \gamma \\ \beta & \theta\beta & \gamma & \cdot \end{pmatrix} \end{array},$$

where once more $Q\underline{1} = \underline{0}$ determines the diagonal elements.

In both of these examples, all the parameters indicated must be positive. The statistical problem is to estimate the parameters of Q_X and Q_Y (and possibly the initial distribution $\underline{s}$ if this is assumed unknown) using the data matrix N, and cell probabilities determined by (1.1), (1.2) and (1.4).

Assuming that sites evolve independently of one another, the data matrix N has the structure of a 16-cell multinomial trials experiment, and we wish to estimate the parameters $\underline{\pi} = (\pi_1, \ldots, \pi_p)$ of the cell probabilities $F = \{f_{ij}(\underline{\pi})\}$. If $\underline{s}$ must be estimated, then $p = 15$ for model (K), and $p = 11$ for model (TK). We chose to estimate the parameters by maximum likelihood (or, equivalently, minimum χ^2) methods. The theory of maximum likelihood estimation in multinomial trials is, of course, well documented; a good review is given by Cox (1984). We also estimated the asymptotic variance-covariance matrix of the estimates for large sample size (= sequence length), n.

The parameter of particular interest here is the mean number K_X (resp., K_Y) of substitutions in the X (resp., Y) species in the interval [0, T]. Recall that if a Markov chain has generator $Q = \{q_{ij}\}$ with $q_i := -q_{ii}$, and initial distribution $\underline{s}$, then the mean number of changes of state in [0, T] is

$$K = \sum_i s_i \sum_j q_j \int_0^T (e^{Qs})_{ij} ds. \qquad (2.1)$$

Notice that if $\underline{s}$ is a stationary distribution for Q, then (2.1) reduces to

$$K^e = T \sum_i s_i q_i, \qquad (2.2)$$

the 'e' denoting equilibrium value. Notice also that if a Markov chain has generator QT, then the mean number of jumps it makes in [0, 1] is given by (2.1), with T = 1, Q = QT. The parameter T is confounded in this estimation problem; from now on, we absorb T into Q, and so take T = 1 in (2.1) and (2.2). In the present context, then, we want to use our estimates of the parameters in Q_X, Q_Y and $\underline{s}$ to estimate K_X and K_Y, and their joint asymptotic distribution.

III. COMPUTATIONAL ASPECTS

The maximum likelihood estimates of the parameters $\underline{\pi}$ of the model may be found using a constrained optimisation package. The constraints involve positivity of the parameters in the two Q-matrices, and, if the initial probabilities $\underline{s}$ are to be estimated, then $0 \le s_1 + s_2 + s_3 \le 1$; $s_1, s_2, s_3 \ge 0$ must hold. Our approach was to transform out the constraints, converting the problem into an unconstrained one, and then use one of the IMSL optimisation algorithms, ZXMIN or ZXSSQ.

From the point of view of asymptotic theory needed here, it seems to be very hard to prove that a unique solution of the likelihood equations exists. Further, the computational work often found solutions in which some elements of the Q-matrices were (algorithmically) zero. The approach adopted here was an exploratory one: We started the algorithm from several initial positions, and compared the results. Any parameters in the Q-matrices that were computationally zero (say, $\le 10^{-6}$) were set to

zero, and not used as parameters. This reduces the dimension p of the problem, and thus increases the degrees of freedom for the goodness of fit test of the fitted model
to the data. All derivatives were calculated using multipoint forward difference formulae (to avoid negative parameter values), as no explicit formulae for such derivatives seem to be useful. The values of the cell probabilities involve calculation of matrix exponentials, $\exp\{Q_X\}$ and $\exp\{Q_Y\}$; recall (1.1) and (1.2). For the model (K), we used the exact results of Gojobori et al. (1982). For model (TK), we used a diagonalisation method, falling back on an efficient series algorithm if this failed.

IV. AN EXAMPLE

The data here are taken from the EMBL sequence library. The sequences are from bovine (X) and mouse (Y) mitochondrial genomes [Anderson et al. (1982), Bibb et al. (1981)], and come from the sequence of third base positions of the genes cytochrome B, cytochrome oxidase I, II and III, and Atpase 6. The base length is n =1601, and the data matrix N is given by

$$N = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} 463 & 91 & 96 & 28 \\ 86 & 140 & 100 & 5 \\ 120 & 164 & 227 & 6 \\ 49 & 14 & 8 & 4 \end{pmatrix} \end{array} .$$

We ran five repetitions of the program, obtaining the same solutions for each run. For the model (K), the estimated Q-matrices were

$$Q_X = \begin{pmatrix} -.262 & .070 & .096 & .096 \\ .0 & -.192 & .096 & .096 \\ .301 & .301 & -.602 & .0 \\ .301 & .301 & .0 & -.602 \end{pmatrix}, \quad Q_Y = \begin{pmatrix} -.209 & .097 & .056 & .056 \\ .0 & -.112 & .056 & .056 \\ .376 & .376 & -.752 & .0 \\ .376 & .376 & .0 & -.752 \end{pmatrix}$$

while for model (TK), the estimated Q-matrices were

$$Q_X = \begin{pmatrix} -.212 & .015 & .087 & .110 \\ .015 & -.212 & .110 & .087 \\ .322 & .406 & -.743 & .015 \\ .406 & .322 & .015 & -.743 \end{pmatrix} \quad Q_Y = \begin{pmatrix} -.111 & .0 & .041 & .070 \\ .0 & -.111 & .070 & .041 \\ .324 & .577 & -.881 & .0 \\ .557 & .324 & .0 & -.881 \end{pmatrix}$$

The estimated base composition of the ancestral sequence was (in order A, T, C, G) for model (K): (.406, .070, .520, .004), while for model (TK): (.351, .039, .604, .006). The estimated average number of substitutions since divergence is given via (2.1) by:

	K_X ± std. error	K_Y ± std. error
Model (K)	.463 ± .032	.441 ± .027
Model (TK)	.401 ± .047	.406 ± .051

The average number of substitutions per site, $K_X + K_Y$, was estimated as:

	$(K_X + Y_Y)$ ± std. error
Model (K)	904 ± .042
Model (TK)	.807 ± .043

For model (K), the goodness-of-fit statistic was $\chi^2 = 18.3$ (6 df.) for model (K), and $\chi^2 = 11.5$ (5 df.), both of which are reasonable. The mean number of substitutions per site as estimated by either method is similar to estimates obtained for models in which assumptions (1.3) hold. For example, the reversible model of Tavare (1985) or Lanave et al. (1984) gave 1.09 ± .13. In this last case, the estimate of the ancestral composition was (.436, .231, .296, .037), which should be contrasted with the estimates from the asymmetric models. Finally, despite the asymmetric shape of N, the estimated Q-matrices are qualitatively similar, and no significant difference can be found between K_X and K_Y. However, the asymptotic substitution rate K^e in (2.3) does differ significantly between the species:

	K_X^e ± std. error	K_Y^e ± std error
Model (K)	.311 ± .042	.217 ± .133
Model (TK)	.326 ± .043	.197 ± .034

A more extensive study of the substitution process among a variety of mitochondrial sequences is being prepared; this will also discuss problems of heterogeneity (due to almalgamating different coding regions, or due to

heterogeneity within a single coding region), the estimation of transition and transversion rates, and the independence of bases assumption used here.

IV CONCLUSIONS

This note has focused on statistical methods for the estimation of substitution rates from pairs of DNA sequences. The analysis has allowed for the observed asymmetry in the data. The estimation methods are computational in nature, rather than the analytic "method-of-moments" approaches usually used. It should be pointed out the the restriction to a small class of models was necessary because in the two-sequence case only 15 degrees of freedom are available. The methods developed here are also useful for analysing multiple-sequence data sets, in which general models of the type (1.2) may be fitted.

REFERENCES

Anderson, S., H. L. DeBruijn, A. R. Coulson, I. C. Eperon, F. Sanger and I. G. Young (1982). Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. J. Mol. Biol., 156, 683-717.

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.

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.

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.

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.

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 "Lectures on mathematics in the life sciences", Vol. 17. Amer. Math. Soc., in press.