J. molec. Evolution 2, 87—90 (1972)

Letter to the Editor

On the Stochastic Model for Estimation of Mutational Distance between Homologous Proteins*

Motoo Kimura and Tomoko Ohta

National Institute of Genetics, Mishima, Japan

Received September 18, 1972

Summary. A set of simple equations is derived which gives the relationship between the observed amino acid differences per 100 codons and the evolutionary distance per 100 codons using Holmquist's stochastic model of molecular evolution.

Key words: Molecular Evolution—Evolutionary Distance Estimation.

Comparing homologous proteins and estimating their "evolutionary distance", that is, the number of mutant substitutions involved, have now become a common practice. The classical method due to Zuckerkandl and Pauling (1965) and recently used by Dickerson (1971) which makes Poisson correction for multiple substitutions, although still useful and handy in estimating amino acid substitutions in evolution, fails to detect some "multiple hits" and "back mutations" in addition to synonymous mutations, especially when nucleotide sites are considered.

To overcome this difficulty, Holmquist (1972a) introduced a stochastic model of evolution. The main assumptions of the model are: the "accepted point mutations" (mutant substitutions) occur spatially at random and in uniform probability over the variable part of the structural gene, and at each site a given nucleotide mutates with equal probabilities to any one of the remaining three. He then presents the relation between the observed amino acid differences and the evolutionary distance per 100 codons in his Table 1 and Fig. 1 under the designation REHs (random evolutionary hits).

The purpose of the present note is to show that these two quantities can be expressed by pair of simple equations connected by an observable parameter λ.

---

* Contribution No. 910 from the National Institute of Genetics, Mishima, Shizuoka-ken 411 Japan.

Let P_d (a.a.) be the fraction of amino acid sites by which two homologous proteins differ from each other, and let D_E be the average number of mutant substitutions per codon that separate the two cistrons of these proteins. Then, we shall show that these quantities are given by a set of equations,

$$P_d(\text{a.a.}) = 1-(1-\lambda)^2\,(1-\tfrac{1}{4}\lambda) \tag{1}$$

and

$$D_E = -\tfrac{9}{4}\ln\,(1-\tfrac{4}{3}\lambda), \tag{2}$$

where λ represents the fraction of nucleotide sites for which the two cistrons (nucleotide sequences) differ from each other ($0 \leqq \lambda \leqq \tfrac{3}{4}$).

Throughout this paper we consider expectations rather than sample values, so we regard P_d(a.a.) as the probability that the two proteins differ at a randomly chosen amino acid site. Likewise, we regard λ as the probability that two sequences differ at a randomly chosen nucleotide site. Then, Eq. (1) can be derived from the consideration that $1-P_d$(a.a.) represents the probability that two homologous codons code for the same amino acid, and this is equal to

$$(1-\lambda)^2\{(1-\lambda)+\tfrac{3}{4}\lambda\},$$

because $(1-\lambda)^2$ represents the probability that the two codons are the same with respect to the first two positions, while $(1-\lambda)$ and $\tfrac{3}{4}\lambda$ in the braces give respectively the probability that the third position is the same and the probability that the third position is different but codes for the same amino acid. The last mentioned probability (i.e. $\tfrac{3}{4}\lambda$) is an approximation and is based on the consideration (as is evident from the code table) that roughly in half the cases a change in the third position leads to a change of amino acid with probability 1/2 (purines vs. pyrimidines).

In order to derive Eq. (2), let K be the average number of nucleotide substitutions per site since the divergence of the two cistrons (nucleotide sequences). Then K is given by Eq. (16) in Holmquist (1972b) as

$$K \equiv \frac{X}{L} = \frac{\ln\left(1-\frac{4}{3}\lambda\right)}{L\ln\left(1-\frac{4}{3L}\right)}. \tag{3}$$

In his terminology, $\lambda = N'(x)/L$. Thus, if the number of nucleotides in the cistron (L) is large, K is given with good approximation by

$$K = -\tfrac{3}{4}\ln\,(1-\tfrac{4}{3}\lambda). \tag{4}$$

An equivalent formula is given by Jukes and Cantor (1969). Since the average number of mutant substitutions per codon is $3K$, which we denote by D_E, Eq. (2) follows immediately from Eq. (4). Also, from Eq. (4), we can derive the large sample variance of K as follows. Let $\delta\lambda$ and δK be,

respectively, small changes in λ and K. Then

$$\delta K = \frac{\delta\lambda}{1-\frac{4}{3}\lambda},$$

so that

$$\sigma_K^2 = E\{(\delta K)^2\} = \frac{E\{(\delta\lambda)^2\}}{(1-\frac{4}{3}\lambda)^2},$$

where E stands for the expectation operator. Then, noting that the sampling variance of λ is $\lambda(1-\lambda)/L$, and substituting this for $E\{(\delta\lambda)^2\}$, we obtain

$$\sigma_K^2 = \frac{\lambda(1-\lambda)}{L(1-\frac{4}{3}\lambda)^2}. \tag{5}$$

This variance may be pertinent when we compare two values of K that are estimated using Eq. (4) from two independent sets of comparisons of nucleotide sequences, and try to judge if these K values are statistically different.

In addition, if our aim is to estimate D_E through a set of Eqs. (1) and (2) by using observed amino acid differences, the estimate has the variance

$$\sigma_{DE}^2 = \frac{16P_d(1-P_d)}{(1-\lambda)^2(3-\lambda)^2(1-\frac{4}{3}\lambda)^2 n_{aa}}, \tag{6}$$

where n_{aa} is the number of amino acid sites per protein.

From the definitions of P_d(a.a.) and D_E in the above treatment, it is evident that $100\,P_d$(a.a.) corresponds to the observed amino acid differences per 100 codons and $100\,D_E$ corresponds to the evolutionary distance per 100 codons in Holmquist's terminology. The following table (Table 1) lists numerically the relation between these two quantities for various values of λ. When we plotted these on Holmquist's Fig. 1, taking $100\,D_E$ as abscissae

Table 1. The relationship between $100\,P_d$(a.a.), the expected amino acid differences in 100 codons, and $100\,D_E$, the expected evolutionary distance per 100 codons, given for various values of the parameter λ

λ	$100\,P_d$(a.a.)	$100\,D_E$
$0.0\dot{3}$	7.3	10.2
$0.0\dot{6}$	14.3	20.9
0.10	21.0	32.2
0.20	39.2	69.7
0.30	54.7	114.8
0.40	67.6	171.3
0.50	78.1	246.9
0.60	86.4	361.7
0.65	89.7	452.8
0.70	92.6	608.6
0.73	94.0	814.6
0.75	94.9	∞

and $100\,P_d$ (a.a.) as ordinates, we have found that the resulting curve is almost indistinguishable from his REHs curve.

We would like to thank Drs. J. Felsenstein and R. Holmquist for reading the first draft and for stimulating discussions.

References

Dickerson, R. E.: J. molec. Evolution **1**, 26 (1971).
Holmquist, R.: J. molec. Evolution **1**, 211 (1972a).
Holmquist, R.: J. molec. Evolution **1**, 115 (1972b).
Jukes, T. H., Cantor, C. R.: Evolution of protein molecules. In: Mammalian protein metabolism. H. N. Munro ed., p. 21–132. New York: Academic Press 1969.
Zuckerkandl, E., Pauling, L.: Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins. V. Bryson, & H. J. Vogel eds., p. 97—166. New York: Academic Press 1965.

Motoo Kimura
Tomoko Ohta
National Institute of Genetics
Yata 1, 111, Mishima
Shizuoka-ken, 411 Japan