

Program for Estimating the Number of Nucleotide Substitutions Using Different Methods

J. M. Comeron

Quantification of genetic evolutionary distances in terms of number of nucleotide substitutions per site between two homologous DNA sequences is fundamental in the study of molecular evolution. Nevertheless, this quantification is not plain, and many different statistical methods have been developed to obtain the estimate of the number of nucleotide substitutions per site (K) in order to make corrections for possible multiple hits at a site in nucleotide sequences. Jukes & Cantor's method (J-C), also called the one-parameter method, was the former and is the simplest one, but it tends to give clear underestimates when there is no equal rate of substitutions between any pair of nucleotides (Jukes and Cantor 1969). Starting from this point, a number of different methods have been proposed to better estimate this K value according to data.

Furthermore, when analyzing coding regions, it is useful to distinguish between nonsynonymous and synonymous substitutions (which cause amino acid and no amino acid replacement, respectively). One method for estimating the number of synonymous (K_s) and nonsynonymous (K_a) nucleotide substitutions per site is Nei and Gojobori's (1986) unweighted method (N-G), which computes the number of synonymous and nonsynonymous nucleotide differences by comparing the two sequences codon by codon. When the two codons to be compared differ by two or three nucleotides, this method gives equal weight to the different possible evolutionary pathways. On the other hand, to compute the number of sites under analysis the method takes into account the

fraction of changes at each position of each codon that would be synonymous and nonsynonymous under the assumption of random nucleotide substitution. The K_s and K_a values are proposed to be obtained by using Jukes and Cantor's method to make corrections for multiple hits at a site. Nevertheless, it seems of interest to apply the different proposed methods in estimating the number of synonymous and nonsynonymous nucleotide substitutions per site.

Here, I present a program (DIVERGEN) for estimating the overall K value, or the K_s and K_a values under the Nei-Gojobori unweighted approach, which gives the different estimates obtained by applying several methods for correcting for multiple hits. The 13 methods applied to the calculation of the number of nucleotide substitutions per site are as follows: (1) Jukes and Cantor's one-parameter method; (2-4) Kimura's two-, three-, and six-parameter methods; (5) Takahata and Kimura's four-parameter method; (6) Tajima and Nei's four-parameter method; (7) Gojobori, Ishii, and Nei's six-parameter method; (8-10) Hasegawa et al.'s, Bulmer's and Tamura's extensions of Kimura's two-parameter method; and (11-13) Tajima's algorithms applied to Jukes and Cantor's, Kimura's two-parameter, and Tajima and Nei's methods (Bulmer 1991; Gojobori et al. 1982; Hasegawa et al. 1985; Kimura 1980, 1981; Tajima 1993; Tajima and Nei 1984; Takahata and Kimura 1981; Tamura 1992).

Since this program allows the use of different correction methods, other than the Jukes and Cantor method, it can be useful for choosing more suitable methods, or just the favorite, for single analysis and for comparison tables comprising a wide range of values by using the same method. At the same time, this program could be used as a tool to test the goodness-of-fit of the different methods (by comparing their estimates with expected values).

This program also allows a sliding win-

dow analysis of any given region and when a complete sequence is under analysis, a region subdivision giving the K_s and K_a values for coding regions and the K values for noncoding regions. The program calculates the frequency of the four nucleotides both for the overall analysis and within the synonymous and nonsynonymous sites. Moreover, the program shows the number and frequency of synonymous and nonsynonymous sites under analysis.

Other features of the program include (a) the input format is an ASCII file with LWL91 format, with or without spaces between triplets; (b) there is no limit to the number of sequences to be compared to each other; and (c) there is a maximum length of 9,999 nucleotides for each sequence. When analyzing coding regions under the Nei and Gojobori approach, in order to obtain K_s and K_a values, there are two more features: (1) evolutionary pathways with intermediate Stop codons are skipped in the multiple pathways analysis, and (2) codons with one or more uncertainties ("*" or "N") are not evaluated. Two examples are supplied with the program, both with the input and output data files. These examples display the different input formats and the two kinds of possible analysis.

The program has been written in QuickBASIC V4.50 (Microsoft), and can be run on any IBM-PC compatible computer. The program and sample files are available from the author by sending a 3.5-inch nonformatted diskette or from the EMBL software data library. Alternatively, DIVERGEN will be supplied on a disk with the submission of \$5.00 (U.S.).

From the Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal, 645. 08071 Barcelona, Spain. I thank J. Rozas and M. Aguadé for many useful comments. J. M. Comeron is supported by an FPI fellowship from the Ministerio de Educación y Ciencia, Spain. This work was supported by DGICYT grant no. PB88-0196 to Montserrat Aguadé, Departament de Genètica, Universitat de Barcelona. Address

address above.

The Journal of Heredity 1994:85(6)

References

Bulmer M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol Biol Evol* 8:868-883.

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

Hasegawa M, Kishino H, and Yano T. 1985. Dating of the

original DNA. *J Mol Evol* 22:160-174.

Jukes TH and Cantor CR. 1969. Mammalian protein metabolism: III. New York: Academic Press.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111-120.

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

Nei M and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418-426.

distance between nucleotide sequences. *Mol Biol Evol* 10:677-688.

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

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

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G-C-content biases. *Mol Biol Evol* 9:678-687.

Received November 2, 1993

Accepted April 12, 1994