

K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals

Josep M. Comeron

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

Received on January 21, 1999; revised on May 11, 1999; accepted on May 12, 1999

Abstract

Summary: *K-Estimator 4.5 is a Windows program that estimates the number of nucleotide substitutions per site (divergence) when comparing two aligned nucleotide sequences, both protein-coding and non-coding. Confidence intervals of the divergence estimates are obtained by Monte Carlo simulation.*

Availability: *The program is available for non-profit use via anonymous ftp at ftp.bio.indiana.edu/molbio/mswin.*

Contact: *jcomeron@midway.uchicago.edu*

The estimation of the number of nucleotide substitutions between two nucleotide sequences is a central subject in the study of molecular evolution. The accurate quantification of such numbers as well as of their confidence intervals directly affects the reliability of many tests broadly applied in evolutionary genetics.

For protein-coding regions, it is interesting to estimate separately the number of nucleotide substitutions that do not provoke a change of amino acid and those that do, called synonymous and non-synonymous substitutions, respectively. The estimation of the number of synonymous (K_s) and non-synonymous (K_a) substitutions per site can be obtained by applying several methods. While the different approaches used to correct for multiple hits at a site tend to give equivalent values of divergence for very closely related sequences, the number of

analyzed sites calculated by the different methods may give biased K_s and K_a estimates across the whole range of divergence levels.

Confidence intervals for divergence estimates are usually based on the expected variances due to sampling assuming a normal, symmetrical distribution (Kimura and Ohta, 1972; Kimura, 1980). Two sources of variance however may make this assumption inaccurate: 1) the number of substitutions can be accepted to be Poisson distributed if we suppose constant substitution rate among sites, and 2) the random distribution along a sequence of the same number of substitutions can give different divergence estimates as a result of multiple hits at a site. The higher the divergence value and/or the smaller the number of sites under analysis, the more skewed to high values the expected distribution of divergence estimates.

K-Estimator is a Windows program written in Visual Basic 5.0 (Microsoft©) and can run on any IBM compatible computer under Windows 95/98 or Windows NT. The program accepts several multiple-sequence formats of already aligned nucleotide sequences (ASCII files): Clustal W, PHYLIP, MSF(PileUp)/GCG, GDE, MEGA, NBRF/PIR and LWL(91) (with or without spaces between codons). There is no program limit to the maximum length or number of sequences to be compared. For both non-coding and coding region sequences, it is possible to analyze particular regions or to obtain results from a sliding window analysis.

Table 1. Number of synonymous substitutions per site (K_s) and confidence intervals

Acc. numbers	Species ^b	No. bp ^c	K_s	C. I. ^a			
				$P = 0.05$		$P = 0.01$	
				Min.	Max.	Min.	Max.
M11739-K01259	m-r	183	0.1607	0.0524	0.3008	0.0341	0.3738
M33974-M33976	m-r	213	0.4392	0.2376	0.6949	0.1890	0.8388
X01838-Y00441	m-r	297	0.6645	0.4105	1.0144	0.3501	1.1787
X53331-D00613	h-m	309	1.0900	0.6940	∞ ^d	0.6014	∞ ^d
U12255-L17022	h-m	1083	0.9357	0.7510	1.1829	0.7052	1.3151
U09607-L32955	h-m	3294	0.7824	0.6915	0.8832	0.6663	0.9136

^a10 000 replicates. ^bm, r, and h indicate mouse, rat, and human, respectively. ^cNumber of base pairs compared. ^d30% of replicates give either $K_s > 5.0$ or inapplicable method to correct for multiple hits at a site.

For non-coding regions, the program can estimate the overall (K) number of nucleotide substitutions per site using several multiple-hits at a site correcting methods: Jukes and Cantor's 1-parameter (Jukes and Cantor, 1969), Kimura's 2-p (Kimura, 1980), Tajima and Nei (Tajima and Nei, 1984), and Tajima's 1-p, 2-p and 4-p (Tajima, 1993).

When coding regions are under analysis, K-Estimator 4.5 applies the method described in Comeron (1995) to estimate K_s and K_a . This method, a modification of the method of Li (1993) and Pamilo and Bianchi (1993) (LPB), better quantifies the actual number of transitions and transversions and reduces stochastic errors (see Comeron, 1995, for details and comparison to previous methods). Three genetic codes can be applied: Universal, Vertebrate mitochondrial, or *Drosophila* mitochondrial. Furthermore, three different options can be applied to restrict the codons that are under analysis: 1) *Maximum one substitution per codon* (analyzes only those codons with no or only one substitution), 2) *No three differences per codon* (removes from the analyses those homologous codons that differ in the three positions), and 3) *Only AAs Substitution* (estimates the K_s analyzing only those homologous codons that code for different amino acid but do not differ at the three positions).

K-Estimator 4.5 obtains the Confidence Intervals (*C.I.*) of divergence estimates (K for non-coding regions, and K_s and K_a for coding regions) by Monte Carlo simulations (Comeron, 1995). Computer simulations take into account the following parameters: 1) divergence value; K , or K_s and K_a , 2) number of nucleotides or codons, 3) the transition : transversion ($\alpha : \beta$) substitution ratio, and 4) the G+C content for non-coding regions, and the amino acid composition and G+C content at the third position of codons for coding regions. When $\alpha : \beta$ is different than that expected under random nucleotide substitution, the substitution pattern is biased accordingly to maintain the original G+C percentage. For all simulations, the number of substitutions applied in each replicate follows a random Poisson-distributed number with a mean equal to the estimated number of substitutions (divergence value \times number of analyzed sites). Substitutions are randomly distributed along the sequence. Since most multiple-hits correcting methods can give slightly biased divergence estimates under some conditions, Monte Carlo simulations using a number of substitutions based on these estimates could give inaccurate *C.I.* caused by a biased divergence average. To solve this putative problem, K-Estimator 4.5 first scans for the optimal number of substitutions that will give the closest divergence average to the analyzed divergence value under the queried conditions, and subsequently it runs the final set

of replicates. Confidence intervals for K_s and K_a estimates are analyzed together and can only be obtained after estimating K_s and K_a with K-Estimator 4.5; the number of codons, the amino acid composition (average of the two compared sequences), the G+C content at the third position of codons, as well as the number of synonymous and non-synonymous substitutions, are fixed from the analyzed sequences.

Confidence intervals are obtained directly from the null distribution of the divergence estimates from each replicate. The program can also calculate the exact probability of obtaining any particular divergence value (K for non-coding regions, and K_s , K_a , and K_a/K_s for coding regions). Table 1 shows the K_s estimates for a few interspecific comparisons of homologous coding sequences of mouse, rat, and human and their confidence intervals obtained by K-Estimator 4.5.

Results of both divergence and confidence intervals for divergence estimates analyses can be printed and/or saved as independent text files. Also, a file with a MEGA-compatible distance matrix format (lower-left matrix) can be obtained for any estimated divergence value.

Acknowledgments

I thank A. Llopart for many suggestions to improve the computer program and valuable discussions on the methods for estimating K_s and K_a . J.M.C. is supported by a postdoctoral fellowship from Ministerio de Educacion y Cultura, Spain.

References

- Comeron, J.M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.*, **41**, 1152–1159.
- Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In Munro, H.W. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–120.
- 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. and Ohta, T. (1972) On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.*, **2**, 87–90.
- Li, W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.
- Pamilo, P. and Bianchi, N.O. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.*, **10**, 271–281.
- Tajima, F. (1993) Unbiased estimation of evolutionary 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.