

# Letter to the Editor

## Approximate Variance of Nucleotide Divergence Between Two Sequences Estimated From Restriction Fragment Data

THERE is an increasing number of studies that make use of DNA fragments to characterize genetic differences among individuals, populations, and species. This is the case, for instance, of RAPDs-PCR and related methods (HADRYN *et al.* 1992), where by means of short oligonucleotides, fragments of genomic DNA are amplified. However, the genetic differences obtained are not treated quantitatively to ascertain genetic distances due to several problems described elsewhere (CLARKE and LANIGAN 1993; LYNCH and MILLIGAN 1994). Essentially, the estimate of nucleotide divergence from RAPDs bands between two genomes is formally similar to the one reported by NEI and LI (1979) when fragments are obtained from restriction analysis (CLARKE and LANIGAN 1993).

NEI and LI (1979) developed a method to estimate the number,  $d$ , of nucleotide substitutions per site between two DNA sequences when restriction fragment data are available for both species. The number  $d$  can be approximated by

$$d = -\frac{2}{r} \ln G, \quad (1)$$

where  $G$  is  $e^{-r\lambda}$ ,  $r$  being the length of the nucleotide sequence recognized by the restriction enzyme used, which is equivalent to the length of the oligonucleotide sequence used to amplify DNA in RAPD-PCR (CLARKE and LANIGAN 1993),  $\lambda$  is the nucleotide substitution rate per site per year, and  $t$  is the time elapsed since the divergence of the two sequences being considered. According to NEI and LI (1979) (see also NEI 1987 p. 106), the estimate  $G$  can be derived by the iterative solution of the relation

$$F = \frac{G^4}{3 - 2G}, \quad (2)$$

where  $F$  is the expected proportion of shared DNA fragments between two sequences and is estimated by

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y}, \quad (3)$$

where  $m_X$  and  $m_Y$  represent the total number of restriction fragments in sequences  $X$  and  $Y$ , respectively, and  $m_{XY}$  represents the number of fragments shared by both sequences (NEI 1987, p. 106). To make statistically valid tests between different values of  $d$ , it is necessary to know its variance. In this letter we derive an approxi-

mate expression for the variance of  $d$  when an estimate of  $\text{Var}(\hat{F})$  is already available (NEI and TAJIMA 1981) and check this expression by numerical simulation.

NEI and TAJIMA (1981) derived the following expression for the sampling variance of  $\hat{F}$ :

$$\text{Var}(\hat{F}) = \frac{1}{\bar{m}} \{ \hat{F}(1 - \hat{F}) - \hat{F}^2(1 - \sqrt{\hat{F}}) [1 + \frac{1}{2}(1 - \sqrt{\hat{F}})] \}, \quad (4)$$

where  $\bar{m}$  is the average number of restriction fragments in the sequences analyzed.

Let us assume that an estimate of  $F$  has been obtained empirically and that the solution of Equation 2 has provided an estimate for  $G$ . The application of the Var operator to both sides of Equation 2 gives

$$\text{Var}(\hat{F}) = \text{Var}\left(\frac{G^4}{3 - 2G}\right). \quad (5)$$

The use of FISHER's approximate variance formula (FISHER 1925), or delta method, retaining up to third order term in Taylor's expansion for the computation of the previous variance leads to the expression

$$\text{Var}(\hat{F}) = \frac{36G^5(2 - G)}{(3 - 2G)^4} [G(2 - G) \text{Var}(G) + \frac{2(2G^2 - 8G + 9)}{3 - 2G} \mu_3(G)], \quad (6)$$

where

$$\mu_3(G) = \frac{(3 - 2G)^6}{216G^9(2 - G)^3} \mu_3(F), \quad (7)$$

in which  $\mu_3(\cdot)$  represents the third order central moment of the corresponding random variable.

The third order moment of  $F$  has been derived as

$$\mu_3(F) = \frac{F(1 - F)(1 - 2F)}{n_0^2} - \frac{F_3 \{ n_0 \sqrt{F}(1 - \sqrt{F})(1 - 2\sqrt{F}) + \sqrt{n_0(1 - \sqrt{F})} \}}{4n_0^3} - \frac{3F^2(1 - n_0\sqrt{F} - 2n_0F + 2n_0^2\sqrt{F^3})}{4n_0^4}. \quad (8)$$

Again, use of FISHER's approximate variance on Expression 1 leads to

$$\text{Var}(\hat{d}) = \frac{4}{r^2 G^2} \left[ \text{Var}(G) - \frac{\mu_3(G)}{G} \right]. \quad (9)$$

The substitution of Expression 9 into Equations 6 and 7, and further rearrangement leads to

$$\text{Var}(\hat{d}) = \frac{(3 - 2G)^4}{9r^2 G^8 (2 - G)^2} \left[ \text{Var}(\hat{F}) - \frac{72 - 117G + 64G^2 - 12G^3}{6G^4 (2 - G^2)} \mu_3(\hat{F}) \right]. \quad (10)$$

To test the accuracy of the preceding derivation and to compare the variance estimates of nucleotide divergence obtained from the analyses of restriction fragments and nucleotide sequences, a computer simulation, following the procedure described in Li (1981), has been carried out.

The simulation analyses were based on three random sequences (1000, 10,000, and 100,000 bp long, respectively) with equal composition frequencies of the four nucleotides. These sequences were made to evolve allowing only for nucleotide substitutions. Each original sequence was changed randomly, giving rise to 2000 pairs of derived sequences for each evolutionary rate. Evolutionary rates,  $2\lambda t$ , varied from 0.002 to 0.05 substitutions per nucleotide and unit time along each branch. The resulting pairs were compared by three different procedures.

First, each pair of sequences was compared using Jukes-Cantor correction for estimating nucleotide divergence. Second, two different restriction digestions were simulated, using either one four-cutter or 10 different six-cutters with recognition sequences obtained at random. Divergence between any two sequences was estimated either from the length of the fragments obtained in both restrictions using Equations 1-3 and from restriction site data, using Equation 5.42 in NEI (1987, p. 101). Variances for the estimated divergence for each pair of sequences were computed according to Equation 5.4 in NEI (1987, p. 66) for nucleotides, Equation 10 for restriction fragments data and Equation 5.45 from NEI (1987, p. 101) for restriction sites data.

Values for the average divergences estimated directly from the sequences and restriction sites and fragments are shown in Table 1. This table also shows the variances of the previous estimates and the average value of the 2000 variances of the estimated distances computed using nucleotide and restriction sites and fragments' divergences for each evolutionary rate.

The reliability of the simulated evolutionary process for all the rates and sequence lengths can be checked by the Jukes-Cantor estimates of nucleotide divergence. Also, the Monte Carlo method for checking the variance expression (5.4) in NEI (1987, p. 66) is corroborated when the second and third columns under the

heading "Restriction Sites" in the table are compared. Taking this into account, the estimation of evolutionary distances from sites and length of restriction fragments shows the following features.

For restriction site data, better estimates are obtained using one single four-cutter than 10 six-cutters. This statement is independent of the length of the sequence used in the simulation, although generally better estimates are obtained using longer sequences. Using one four-cutter, and for  $L = 1000$  or  $L = 10000$ , generally a slight overestimate of  $\text{Var}(d)$  is obtained using Equation 5.45 from NEI (1987). For  $L = 100,000$  and all rates and for two previous  $\text{Var}(d)$  is usually underestimated by this equation. When ten six-cutters are used, the estimated distances are only acceptable for  $L = 100,000$  and are clear underestimates for smaller lengths. These underestimates cannot be explained by a wrong simulation as deduced in the previous paragraph. There is no clear pattern for the corresponding variances, where small under- and overestimates can be observed for all lengths and evolutionary rates.

The use of restriction fragment lengths for estimating sequence divergence presents several problems. First, nucleotide divergences can only be estimated reliably from a combination of the right sequence length and number of restriction fragments. In our case, this was achieved using either  $L = 10,000$  and one four-cutter or  $L = 100,000$  and 10 six-cutters. All other combinations resulted in serious underestimates of the simulated divergence. For small sequence lengths or numbers of generated fragments, this could be due to the small number of fragments generated. For larger sequences, the explanation rests on the redundancy of nonhomologous fragments with equal length generated when  $2\lambda t \geq 0.02$ . For the two cases with acceptable estimates of divergence ( $L = 10,000$ ,  $r = 4$  and  $L = 100,000$ ,  $r = 6$ ), the variance estimates obtained using Equation 10 always underestimate the variances deduced from Monte Carlo replicates. This leads to an increased type I error probability if it is accepted that the variances obtained from the 2000 estimates of  $d$  are a good approximation to the actual variances.

There are two serious limitations to the use of restriction fragment data for the estimation of nucleotide divergences: the range of divergences to which the method can be reliably applied is very limited ( $2\lambda t \leq 0.05$ ) and the lack of an expression for the variance of the estimated divergence. We have tried to provide a solution for this last problem by deriving Equation 10 and checking its accuracy by Monte Carlo simulation. Although the variances obtained from the simulations and by direct application of Equation 10 are not as coincident as desirable, nevertheless a quite good approximation to the magnitude of the variance can be obtained from the expression here proposed. The values derived from Equation 10 can be used for per-

TABLE 1  
Estimates of divergence

$2\lambda t$	Length	Nucleotide sequences						Restriction sites						Restriction fragments					
		4-cutter		6-cutter		4-cutter		6-cutter		4-cutter		6-cutter		4-cutter		6-cutter			
		$d$	sim	eq	$d$	sim	eq	$d$	sim	eq	$d$	sim	eq	$d$	sim	eq	$d$	sim	eq
0.002	1000	0.0021	1.970E-6	2.06E-6	0.0019	4.086E-5	4.507E-5	0.0003	2.329E-6	2.302E-6	0.0022	5.149E-5	2.840E-4	0.0003	2.680E-6	3.977E-5			
	10,000	0.0020	2.080E-7	2.002E-7	0.0020	6.331E-6	6.165E-6	0.0014	3.094E-6	3.246E-6	0.0020	6.229E-6	8.913E-6	0.0014	3.281E-6	5.234E-6			
	100,000	0.0020	2.047E-8	2.000E-8	0.0020	6.604E-7	6.219E-7	0.0020	7.074E-7	6.638E-7	0.0019	5.930E-7	2.818E-7	0.0020	7.150E-7	3.161E-7			
0.005	1000	0.0050	5.157E-6	5.02E-6	0.0047	2.156E-4	2.345E-4	0.0011	7.832E-6	7.776E-6	0.0061	3.663E-4	8.837E-4	0.0012	9.103E-6	3.200E-5			
	10,000	0.0050	4.938E-7	5.060E-7	0.0045	1.180E-5	1.195E-5	0.0034	7.079E-6	6.610E-6	0.0046	1.205E-5	9.424E-6	0.0035	7.351E-6	5.133E-6			
	100,000	0.0050	4.642E-8	5.034E-8	0.0050	1.848E-6	1.635E-6	0.0048	1.462E-6	1.572E-6	0.0048	1.601E-6	6.301E-7	0.0048	1.498E-6	6.515E-7			
0.01	1000	0.0100	1.017E-5	1.013E-5	0.0093	3.378E-4	3.569E-4	0.0022	1.587E-5	1.626E-5	0.011	4.920E-4	6.912E-4	0.0024	1.862E-5	4.043E-5			
	10,000	0.0100	1.007E-6	1.012E-6	0.0092	2.743E-5	2.690E-5	0.0067	1.460E-5	1.490E-5	0.0091	2.736E-5	1.631E-5	0.0068	1.545E-5	9.338E-6			
	100,000	0.0100	1.021E-7	1.013E-7	0.0096	2.815E-6	3.024E-6	0.0094	3.188E-6	3.155E-6	0.0090	2.328E-6	1.113E-6	0.0094	3.280E-6	1.307E-6			
0.02	1000	0.0200	2.034E-5	2.042E-5	0.0169	5.162E-4	5.323E-4	0.0046	2.869E-5	2.994E-5	0.0195	7.424E-4	6.300E-4	0.0050	3.596E-5	3.944E-5			
	10,000	0.0200	2.187E-6	2.044E-6	0.0184	5.804E-5	5.491E-5	0.0135	2.839E-5	3.052E-5	0.0182	5.656E-5	2.866E-5	0.0137	3.176E-5	1.684E-5			
	100,000	0.0201	2.128E-7	2.052E-7	0.0208	7.031E-6	7.898E-6	0.0192	7.251E-6	7.109E-6	0.0185	5.308E-6	2.915E-6	0.0191	7.500E-6	3.250E-6			
0.03	1000	0.0300	3.116E-5	3.109E-5	0.0229	5.742E-4	6.103E-4	0.0057	3.398E-5	3.949E-5	0.0262	8.386E-4	5.946E-4	0.0060	3.936E-5	4.736E-5			
	10,000	0.0301	3.270E-6	3.104E-6	0.0368	1.823E-4	1.945E-4	0.0205	4.827E-5	5.076E-5	0.0366	1.983E-4	1.140E-4	0.0211	6.066E-5	2.915E-5			
	100,000	0.0303	3.014E-7	3.119E-7	0.0292	1.061E-5	1.002E-5	0.0287	1.310E-5	1.111E-5	0.0240	6.396E-6	3.518E-6	0.0284	1.427E-5	5.666E-6			
0.04	1000	0.0400	4.108E-5	4.197E-5	0.0300	9.923E-4	1.045E-3	0.0046	3.476E-5	3.819E-5	0.0346	1.362E-3	8.081E-4	0.0050	4.255E-5	6.295E-5			
	10,000	0.0401	4.181E-6	4.174E-6	0.0404	1.721E-4	1.862E-4	0.0291	9.629E-5	1.021E-4	0.0394	1.647E-4	8.999E-5	0.0303	1.294E-4	6.518E-5			
	100,000	0.0405	4.298E-7	4.215E-7	0.0411	1.703E-5	1.631E-5	0.0388	1.641E-5	1.657E-5	0.0323	9.419E-6	5.796E-6	0.0383	1.831E-5	9.375E-6			
0.05	1000	0.051	5.336E-5	5.344E-5	0.0385	9.265E-4	9.477E-4	0.0069	5.071E-5	5.448E-5	0.0410	1.189E-3	6.484E-4	0.0075	6.440E-5	6.542E-5			
	10,000	0.0503	5.011E-6	5.293E-6	0.0489	1.985E-4	1.992E-4	0.0340	9.846E-5	1.012E-4	0.0480	2.216E-4	1.156E-4	0.0347	1.268E-4	6.373E-5			
	100,000	0.0507	5.262E-7	5.339E-7	0.0495	1.943E-5	1.874E-5	0.0466	1.906E-5	1.923E-5	0.0361	9.510E-6	6.260E-6	0.0459	2.345E-5	1.180E-5			

$d$ , average value of 2000 comparisons for each rate of evolution ( $2\lambda t$ ), sequence length, and method of estimation, which follows: divergence from nucleotide sequences using Jukes-Cantor's correction, nucleotide divergence; divergence from restriction site data generated by digestion with one 4-cutter restriction endonuclease, 4-cutter, or 10 6-cutter restriction endonucleases, 6-cutter; divergence from restriction fragment lengths from the two previously described digestions. In all cases variances of the corresponding divergences were obtained directly from the Monte-Carlo replicates (sim) and from the corresponding equations as described in the text (eq).

forming conservative tests among pairs of sequences being compared. The increasing use of the RAPD-PCR method for deriving phylogenetic relationships, especially at the intraspecific and intrageneric levels, could benefit from the use of this expression for providing approximate tests of the derived relationships.

We have tried to test the accuracy of our Monte Carlo simulations by comparing the divergences obtained after evolution of one single sequence into pairs of derived ones using three different methods: nucleotide, restriction site, and restriction fragment data. We agree with COCKERHAM and WEIR (1993) in that simulation results should be thoroughly checked against analytically derived expressions. This has led us to identify some possible pitfalls in the use of simulations in sequence evolution and later comparing sequences by sampling a subset of their nucleotides, either by restriction sites or fragments. Even if the simulated and analytically deduced results are not coincident for restriction fragment data, the two other estimations allow us to be confident on the reliability of our simulations.

We are indebted to Drs. J. FERRÁNDIZ and M. SENDRA for valuable suggestions and comments and to Dr. W.-H. Li for his support and suggestions with previous versions of the manuscript. S.F.E. has been supported by a fellowship from Conselleria d'Educació i Ciència, Generalitat Valenciana (Spain). This work has been funded by grants PB93-0690 and PB93-0350 from DGICYT (Spain).

FERNANDO GONZÁLEZ-CANDELAS,  
SANTIAGO F. ELENA and ANDRÉS MOYA  
Departament de Genètica and  
Servei de Bioinformàtica  
Facultat de Biologia  
Universitat de València Estudi General  
Dr. Moliner, 50  
E-46100 Burjassot  
València, Spain

#### LITERATURE CITED

- CLARKE, A., and C. M. S. LANIGAN, 1993. Prospects for estimating nucleotide divergence with RAPDs. *Mol. Biol. Evol.* 10: 1096-1111.
- COCKERHAM, C. C., and B. S. WEIR 1993. Estimation of gene flow from F-statistics. *Evolution* 47: 855-863.
- FISHER, R. A., 1925. *Statistical Methods for Research Workers*, Ed. 13, Hafner, New York.
- HADRYS, H., M. BALICK and B. SCHIERCRATER, 1992. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Mol. Ecol.* 1: 55-63.
- LI, W.-H., 1981. A simulation study on Nei and Li's model for estimating DNA divergence from restriction enzyme maps. *J. Mol. Evol.* 17: 251-255.
- LYNCH, M., and B. G. MILLIGAN, 1994. Analyzing population genetic structure with RAPD markers. *Mol. Ecol.* 3: 91-100.
- NEI, M., 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- NEI, M., and W.-H. LI, 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* 76: 5269-5273.
- NEI, M., and F. TAJIMA, 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics* 97: 145-163.