Genet. Res., Camb. (1997), **69**, *pp.* 45–48 *With* 1 *text-figure*

Using the variance of pairwise differences to estimate the recombination rate

JOHN WAKELEY*
Department of Biological Sciences, Rutgers University, New Jersey, USA

(*Received 12 August 1996 and in revised form 14 October 1996*)

Summary

A new estimator is proposed for the parameter $C = 4Nc$, where N is the population size and c is the recombination rate in a finite population model without selection. The estimator is an improved version of Hudson's (1987) estimator, which takes advantage of some recent theoretical developments. The improvement is slight, but the smaller bias and standard error of the new estimator support its use. The variance of the average number of pairwise differences is also derived, and is important in the formulation of the new estimator.

1. Introduction

Under the neutral theory of molecular evolution, the average number of pairwise nucleotide differences among sequences in a random sample is independent of the recombination rate. If only non-identical pairs are considered, the expectation of this average is equal to $\theta = 4Nu$, where N is the effective population size and u is the neutral mutation rate in an infinite-sites model. The variance of pairwise differences, however, does depend on the recombination parameter, $C = 4Nc$, where c is the recombination rate. Nearly a decade ago, Hudson (1987) made use of this fact and introduced an estimator of C based on the sample distribution of pairwise differences. Since that time, Hudson's estimator has become the most frequently used of the available estimators. Some recent, related theoretical results now suggest improvements to Hudson's original work.

Here, a new estimator of C is proposed which differs from Hudson's in that only non-identical pairs of sequences are considered and because an unbiased estimator of θ^2 is employed in its calculation. The statistical properties of the new estimator are investigated using computer simulations, and are compared with those of Hudson's estimator. The new estimator is less biased and has a smaller standard error.

* Correspondence to: J. Wakeley, Nelson Biological Labs, PO Box 1059, Busch Campus, Piscataway, NJ 08855-1059, USA. Fax: +1-908-445-5870. e-mail: jwakeley@rci.rutgers.edu.

2. The estimators

From a sample of n sequences we can calculate two different averages of the numbers of pairwise differences, which differ according to how many pairwise comparisons are considered. If k_{ij} is the number of differences between two sequences, i and j, these are

$$\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij} \tag{1}$$

and

$$\bar{k} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij}. \tag{2}$$

Thus, π is computed using only non-identical pairs, whereas $\bar{k}$ counts each of these twice and includes the n zero values obtained when each sequence is compared with itself. In a population of constant size with neutral, infinite sites mutation, the expectation of π is θ (Watterson, 1975; Tajima, 1983). Since (2) can be rewritten as $\pi(n-1)/n$, the expectation of $\bar{k}$ is equal to $\theta(n-1)/n$.

Corresponding to (1) and (2), two variances can also be calculated:

$$S_\pi^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (k_{ij} - \pi)^2, \tag{3}$$

$$S_k^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (k_{ij} - \bar{k})^2. \tag{4}$$

When there is no recombination, the expectation of S_π^2 is given by

$$\mathrm{E}(S_\pi^2) = \left[\frac{2(n-2)}{3(n-1)}\right]\theta + \left[\frac{(7n+3)(n-2)}{9n(n-1)}\right]\theta^2 \tag{5}$$

https://doi.org/10.1017/S0016672396002571 Published online by Cambridge University Press

(Wakeley, 1996). Since (4) can be rewritten as

$$S_k^2 = \left(\frac{n-1}{n}\right) S_\pi^2 + \left(\frac{n-1}{n^2}\right)\pi^2, \tag{6}$$

the expectation of S_k^2 becomes

$$\mathrm{E}(S_k^2) = \left[\frac{(n-1)(2n-1)}{3n^2}\right]\theta + \left[\frac{(n-1)(7n^2+7n-6)}{9n^3}\right]\theta^2. \tag{7}$$

Equation (7) follows from the substitution into (6) of expression (5) and the expression for $\mathrm{E}(\pi^2)$ employed by Tajima (1993) to develop an unbiased estimator of the variance of π when there is no recombination.

Hudson (1987) derived the expectation of S_k^2 when there is recombination. His expression can be written

$$\mathrm{E}(S_k^2) = \left[\frac{(n-1)(2n-1)}{3n^2}\right]\theta + g_k(C,n)\,\theta^2. \tag{8}$$

The expression for $g_k(C,n)$ is reproduced in the Appendix in a different format from that of Hudson (1987). The limit of $g_k(C,n)$ as C approaches zero is, then, equal to the term multiplying θ^2 in (7). Hudson (1987) proposed the estimator, here called $\hat{C}_k$, that solves

$$S_k^2 = \sum h_j - \sum h_j^2 + g_k(C,n)\left(\frac{n}{n-1}\sum h_j\right)^2, \tag{9}$$

where h_j is the heterozygosity at site j in a sample of DNA sequences. Thus, Hudson's estimator involves using $\sum h_j - \sum h_j^2$ to estimate the first term on the right-hand side of (7) and $[\sum h_j n/(n-1)]^2$ to estimate θ^2, then solving for the value of C that equates the expectation of S_k^2 most closely to its observed value.

The variance of π with recombination can also be obtained. From (6),

$$\mathrm{E}(S_k^2) = \left(\frac{n-1}{n}\right) E(S_\pi^2) + \left(\frac{n-1}{n^2}\right) E(\pi^2), \tag{10}$$

and since $\mathrm{E}(S_\pi^2) = \mathrm{Var}(k_{ij}) - \mathrm{Var}(\pi)$ (Wakeley, 1996),

$$\mathrm{Var}(\pi) = \left(\frac{n}{n-1}\right)\mathrm{Var}(k_{ij}) - \left(\frac{n}{n-1}\right)^2 \mathrm{E}(S_k^2) + \left(\frac{1}{n-1}\right)\mathrm{E}(\pi)^2. \tag{11}$$

Hudson (1983) derived an expression for $\mathrm{Var}(k_{ij})$, given explicitly by Hudson (1990); Hudson (1987) developed $\mathrm{E}(S_k^2)$, reproduced here as (8); and Watterson (1975) gave the familiar result that $\mathrm{E}(\pi) = \theta$. Then, $\mathrm{Var}(\pi)$ with recombination becomes

$$\mathrm{Var}(\pi) = \left[\frac{n+1}{3(n-1)}\right]\theta + f(C,n)\,\theta^2, \tag{12}$$

where $f(C,n)$ is given in the Appendix. As C decreases to zero, (12) approaches Tajima's (1983) result. Expression (12) was also recently derived by Pluzhnikov & Donelly (1996), but for other purposes.

It follows, after some simplification, that

$$\mathrm{E}(S_\pi^2) = \left[\frac{2(n-2)}{3(n-1)}\right]\theta + g_\pi(C,n)\,\theta^2, \tag{13}$$

where $g_\pi(C,n)$ is given in the Appendix, is the expectation of (3) when there is recombination. Accordingly, as C approaches zero, (13) approaches (5).

Tajima (1993) noted that π^2 is a biased estimator of θ^2. Of course, this is true also of Hudson's (1987) estimator, $[\sum h_j n/(n-1)]^2$, since expression (1) is identical to $\sum h_j n/(n-1)$. Expression (12) can be used to obtain an unbiased estimator of θ^2:

$$\widehat{\theta^2} = \frac{p^2 - [(n+1)/3(n-1)]\,\pi}{f(C,n)+1}. \tag{14}$$

Thus, the new estimator of C proposed here solves

$$S_\pi^2 = \left[\frac{2(n-2)}{3(n-1)}\right]\pi + g_\pi(C,n)\left[\frac{\pi^2 - [(n+1)/3(n-1)]\,\pi}{f(C,n)+1}\right], \tag{15}$$

where π and S_π^2 are observed values, calculated from a sample of DNA sequences using (1) and (3). This estimator, called $\hat{C}_\pi$, differs from Hudson's (1987) estimator, $\hat{C}_k$, in two main respects: only the $n(n-1)/2$ unique pairwise comparisons among the n sequences are made, and an unbiased estimate of θ^2 is employed.

3. Performance in simulations

Monte Carlo simulations, using the method of Hudson (1983), were done to assess the statistical properties of $\hat{C}_\pi$, and to compare its performance with that of $\hat{C}_k$. Figure 1 compares estimates of the distributions of $\hat{C}_\pi/C$ and $\hat{C}_k/C$, where C is the true value of the recombination parameter, for the same values of n, C and θ used in figure 2 of Hudson (1987). Arrows used to indicate the means of the estimated distributions show that $\hat{C}_\pi$ is less biased than $\hat{C}_k$. As n, C and θ increase, the performances of the two estimators become more and more similar. Not only is $\hat{C}_\pi$ less biased, its variance is smaller. For the three estimated distributions in Fig. 1, $\mathrm{Var}(\hat{C}_\pi/C)$ equals (*a*) 0·09, (*b*) 1·16 and (*c*) 1·82, and $\mathrm{Var}(\hat{C}_k/C)$ equals (*a*) 0·10, (*b*) 1·24 and (*c*) 2·24. In addition, when $n = 11$ and $C = \theta = 25$ (*c*), 4·7% of the distribution of $\hat{C}_k/C$ lies above 4, compared with 3·2% for $\hat{C}_\pi/C$.

4. An application

Schaeffer & Miller (1993) used $\hat{C}_k$ to estimate the recombination rate in 99 sequences of a $\sim$3·5 kb region containing the alcohol dehydrogenase gene of *Drosophila pseudoobscura*. Their data give 315 as an estimate of C when $\hat{C}_k$ is used, and 282 using $\hat{C}_\pi$. Because of the assumption of infinite-sites mutation, sites showing direct evidence of multiple mutations, i.e. the 27 sites segregating more than two nucleotides,

https://doi.org/10.1017/S0016672396002571 Published online by Cambridge University Press

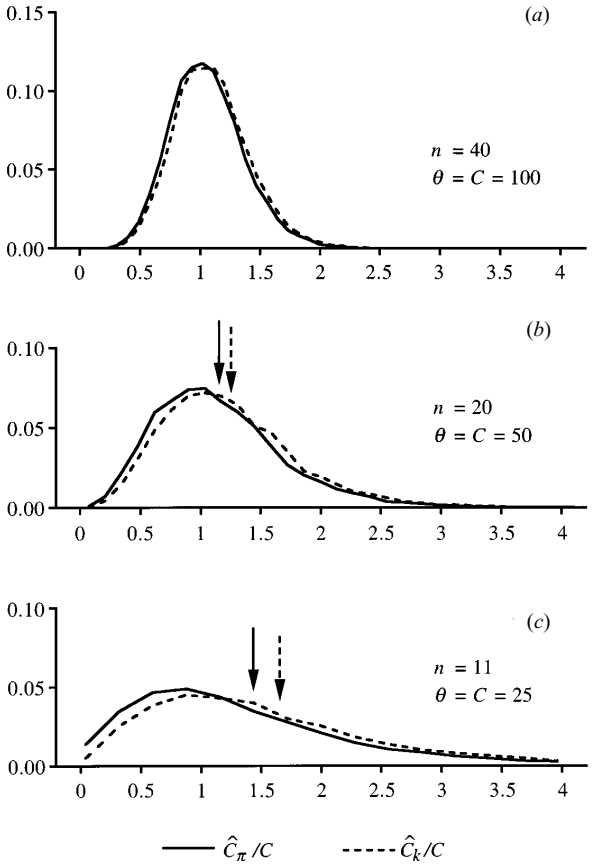

Fig. 1. The estimated distributions of $\hat{C}_\pi/C$ and $\hat{C}_k/C$ for three different sample sizes and values of θ and C. Each distribution is based on 10000 independent samples. The means of the distributions are indicated with arrows. For (*a*) the means are too similar to be distinguished this way; they are 1·06 and 1·10 for $\hat{C}_\pi/C$ and $\hat{C}_k/C$, respectively.

were ignored in these analyses. In addition, Hudson's (1987) simulation method of constructing an approximate 95% confidence interval gives [185, 484] for $\hat{C}_k$, and [172, 453] for $\hat{C}_\pi$. Ten thousand replicates per value of C were used to determine the lower bounds and 5000 to determine the upper bounds for these data. These results are consistent with the simulation results reported above and in Fig. 1: $\hat{C}_\pi$ is smaller, and presumably less biased, than $\hat{C}_k$, and the error of $\hat{C}_\pi$ may be somewhat less than that of $\hat{C}_k$.

5. Discussion

All the main conclusions of Hudson (1987) about $\hat{C}_k$ hold also for $\hat{C}_\pi$, namely the performance of both estimators approaches a satisfactory level only for very large data sets, i.e. like those of the top panel of Fig. 1 with $n = 40$ and $\theta = C = 100$. Since Hudson's (1987) original work, a few such data sets have been generated. The data of Schaeffer & Miller (1993), analysed above, are one example. Fig. 1 shows that it may be better to use $\hat{C}_\pi$ when data are less plentiful. This is because $\hat{C}_\pi$ is less biased than $\hat{C}_k$, and has a smaller standard error. However, for smaller data sets neither of these estimators is expected to be very accurate, so better estimators should be developed. To this end, Hey & Wakeley (1996) have recently developed another estimator of C.

The theory presented above is valuable in one other respect, and that is in quantifying error when π is used to estimate θ. Tajima (1993) gives unbiased estimators of the variance of π under two conditions: no recombination and complete independence of sites. An unbiased estimate of the variance of π under intermediate levels of recombination can be derived from (12) and (14):

$$\widehat{\mathrm{Var}}(\pi) = \frac{(n+1)}{3(n-1)}\pi + f(C,n)\left[\frac{\pi^2 - [(n+1)/3(n-1)]\pi}{f(C,n)+1}\right]. \tag{16}$$

For Schaeffer & Miller's (1993) data, $\pi = 31{\cdot}7$ and $C = 282$, estimated using $\hat{C}_\pi$, so $\widehat{\mathrm{Var}}(\pi)$ becomes 16·1. This can be compared with $\widehat{\mathrm{Var}}(\pi)$ estimated from Tajima's (1993) formulas: for complete linkage, $\widehat{\mathrm{Var}}(\pi) = 194{\cdot}5$, and under free recombination, $\widehat{\mathrm{Var}}(\pi) = 10{\cdot}8$. Because recombination is so frequent in these sequences, the error of this estimate of θ is very small.

Appendix. The portions of $\mathrm{E}(S_k^2)$, $\mathrm{E}(S_\pi^2)$, and $\mathrm{Var}(\pi)$ that depend on C

$$g_k(C,n) = \frac{(n-1)}{n^3 C^2}\left\{-C(2n^2 - nC - 4) - [n^2 - 4n + 14 - C(n^2-2)]\, I_1 + [49n^2 - 52n + 110 + C(15n^2 - 8n + 2)]\frac{I_2}{\sqrt{97}}\right\}$$

$$g_\pi(C,n) = \frac{(n-2)}{n(n-1)\, C^2}\left\{-2C(n+1) - [n - 7 - C(n+1)]\, I_1 + [49n - 55 + C(15n-1)]\frac{I_2}{\sqrt{97}}\right\}$$

$$f(C,n) = \frac{2}{n(n-1)\, C^2}\left\{-2C - [2n(n+1) - 7 - C]\, I_1 + [2n(n+1)(13+2C) - 55 - C]\frac{I_2}{\sqrt{97}}\right\}$$

where

$$I_1 = \log\left[\frac{C^2 + 13C + 18}{18}\right]$$

$$I_2 = \log\left[\frac{(13 - \sqrt{97} + 2C)(13 + \sqrt{97})}{(13 + \sqrt{97} + 2C)(13 - \sqrt{97})}\right]$$

https://doi.org/10.1017/S0016672396002571 Published online by Cambridge University Press

I thank Dick Hudson very much for helpful comments and for making his programs available. I also thank Steve Schaeffer for supplying the DNA data for use in this project. This work was funded by PHS GM 17745-01 from the NIH.

References

Hey, J. & Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics*, in press.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research*, *Cambridge* **50**, 245–250.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (ed. D. J. Futuyma & J. Antonovics), vol. 7. Oxford: Oxford University Press.

Pluzhnikov, A. & Donelly, P. (1996). Optimal sequencing for surveying molecular genetic diversity. *Genetics*, 144: 1247–1262.

Schaeffer, S. W. & Miller, E. L. (1993). Estimation of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **135**, 541–552.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tajima, F. (1993). Measurement of DNA polymorphism. In *Mechanisms of Molecular Evolution* (ed. N. Takahata & A. G. Clark). Sunderland, Mass.: Sinauer Associates.

Wakeley, J. (1996). Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology* **49**, 369–386.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.