STATISTICAL PROPERTIES OF THE NUMBER OF RECOMBINATION EVENTS IN THE HISTORY OF A SAMPLE OF DNA SEQUENCES

RICHARD R. HUDSON AND NORMAN L. KAPLAN

National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Manuscript received November 8, 1985
Revised copy accepted April 30, 1985

ABSTRACT

Some statistical properties of samples of DNA sequences are studied under an infinite-site neutral model with recombination. The two quantities of interest are R, the number of recombination events in the history of a sample of sequences, and R_M, the number of recombination events that can be parsimoniously inferred from a sample of sequences. Formulas are derived for the mean and variance of R. In contrast to R, R_M can be determined from the sample. Since no formulas are known for the mean and variance of R_M, they are estimated with Monte Carlo simulations. It is found that R_M is often much less than R, therefore, the number of recombination events may be greatly underestimated in a parsimonious reconstruction of the history of a sample. The statistic R_M can be used to estimate the product of the recombination rate and the population size or, if the recombination rate is known, to estimate the population size. To illustrate this, DNA sequences from the *Adh* region of *Drosophila melanogaster* are used to estimate the effective population size of this species.

THE neutral infinite-site model introduced by KIMURA (1971) is a natural framework for analyzing nucleotide sequence data. Much of the analytical development of this model has been for the cases in which the rate of recombination is zero or infinite (WATTERSON 1975; EWENS 1979). Recently, HUDSON (1983b) has studied some properties of this model when the rate of intragenic recombination is finite. Using a result of GRIFFITHS (1981) for a two-locus model with finite recombination, he derived a formula for the variance of the number of segregating sites in a sample of size 2 and obtained an approximation for the expected homozygosity.

HUDSON (1983b) has also developed an efficient method for simulating samples from the neutral infinite-site model with finite recombination. His method generates the "history" of a sample. The history of a sample is a collection of correlated family trees, one for each site (for DNA sequence data, each nucleotide is considered a site). The family tree for a site traces the genealogy of a site back to its most recent common ancestor indicating which sampled gametes are most closely related and when the most recent common ancestors occurred. If the rate of recombination is zero, then each site has the same family tree and therefore the history of the sample consists of just one tree. The method for generating this tree depends on results of WATTERSON (1975) (for details see HUDSON 1983a; TAJIMA 1983). On the other hand, if the recombination rate is infinite, then all of the family trees are independent of each other, and each family tree is generated in the same way as when the recombination rate is zero. If the recombination rate is finite, then the topologies and lengths of the branches of the family trees are correlated because of linkage, and generating them is more complex but still possible (HUDSON 1983b).

Let generation t ($t \geq 0$) denote the population t generations before the present one from which the sample is taken. Those gametes in generation t that have descendants in the sample are referred to as the ancestral gametes in generation t. For any site the number of ancestral gametes in generation t is just the number of branches t generations before the present one in the family tree of that site. If an ancestral gamete in generation $t - 1$ is the recombinant descendant of two ancestral gametes in generation t, then we say that a recombination event has occurred in generation t. Let R denote the total number of recombination events in the history of the sample. The object of this paper is to study the statistical properties of R. If the rate of recombination is zero, then $R = 0$, and if the rate of recombination is infinite, then $R = \infty$. Thus, only when the recombination rate is finite and nonzero is R interesting. For this case formulas are derived for the mean and variance of R for arbitrary sample size.

Although R is a quantity of interest from a theoretical point of view, its drawback is that it cannot be evaluated from data, since the history of a sample is never observed. A way of inferring that between two sites at least one recombination event took place in the history of the sample is to use the "four-gamete" test. This test can be explained in the following way. For the infinite-site model the mutation rate for any site is infinitesimal; therefore, at most one mutation event can occur in the history of the sample at that site. Thus, for any two sites there are at most four gametic types in the population. Furthermore, since the model does not allow for back mutation and recurrent mutation, the only way for all four gametic types to be in the sample is for at least one recombination event to have occurred in the history of the sample between the two sites.

Not all recombination events in the history of the sample are revealed by the four-gamete test. For a recombination event to be detected by this test, the history of sampled gametes must have a specific structure and mutations must occur on appropriate lineages of the family trees. It is shown that even for extremely high mutation rates and moderate sample sizes a substantial fraction of the recombination events in the history of the sample can never be detected using the four-gamete test.

Let R_M denote the minimum number of recombination events implied by the data using the four-gamete test (see APPENDIX 2). The statistical properties of R_M are of interest since this quantity arises naturally when one attempts to actually construct the history of the sample. Furthermore, R_M may be useful in estimating the rate of recombination. The mean and variance of R_M are complicated functions of the mutation and recombination rates. It is shown that $E(R_M)$ is an increasing function of the mutation rate and the limiting value of $E(R_M)$ is identified.

The statistical properties of R and R_M for different rates of mutation and recombination are also examined using the simulation methods of HUDSON (1983b). Finally, the results in this paper are discussed in light of the recent data set published by KREITMAN (1983). In particular, the effective population size of *Drosophila melanogaster* is estimated from R_M.

STATISTICAL PROPERTIES OF R

Let $2N$ denote the population size which is assumed to be fixed, c the rate of recombination per generation per gamete, u the rate of mutation per generation per gamete and n the sample size. Both c and u are assumed to be of order $1/N$; therefore, it is convenient to define $\theta = 4Nu$ and $C = 4Nc$. For simplicity the chromosome under study is represented by the interval $[0, 1]$.

Suppose that for any integer, m, the genome is divided into m equal segments which are labeled from 1 to m starting from the left. One then has the identity

$$R = \sum_{i=1}^{m} R_i, \tag{1}$$

where R_i is the number of recombination events in the history of the sample for segment i. Since recombination is assumed to occur uniformly across the genome, all of the R_i have the same distribution. Furthermore, it is shown in APPENDIX 1 that

$$P(R_1 = 0) = \prod_{i=2}^{n} \left(\frac{i-1}{\frac{C}{m} + i - 1} \right), \tag{2}$$

$$P(R_1 = 1) = 1 - P(R_1 = 0) + 0(m^{-2}) \tag{3}$$

and

$$\sum_{j \geq 2} jP(R_1 = j) = 0(m^{-2}). \tag{4}$$

It follows that

$$\begin{aligned} E(R) &= E\left(\sum_{i=1}^{m} R_i\right) \\ &= m\left(1 - \prod_{i=2}^{n}\left(\frac{i-1}{\frac{C}{m} + i - 1}\right) + 0(m^{-2})\right) \\ &= C\left(\sum_{i=1}^{n-1} \frac{1}{i}\right) + 0(m^{-1}). \end{aligned}$$

Letting $m \longrightarrow \infty$, we obtain

$$E(R) = C\left(\sum_{i=1}^{n-1} \frac{1}{i}\right). \tag{5}$$

We next evaluate the variance of R. It follows from (1) that

$$\mathrm{Var}(R) = \sum_{i=1}^{m} \mathrm{Var}(R_i) + 2 \sum_{i<j} \mathrm{cov}(R_i, R_j). \tag{6}$$

In view of (2), (3) and (4)

$$\mathrm{Var}(R_i) = \frac{C}{m}\left(\sum_{i=1}^{n-1} \frac{1}{i}\right) + 0(m^{-2}). \tag{7}$$

The derivation of the formula for $\mathrm{cov}(R_i, R_j)$ requires some additional notation.

For the family tree of the portion of the $(i + 1)$th segment contiguous to the ith segment, let $l(t)$ denote the number of ancestral gametes in generation t, and set $\underset{\rightarrow}{l} = (l(0), l(1), l(2), \ldots)$. In a similar way define $\underset{\rightarrow}{r} = (r(0), r(1), r(2), \ldots)$ for the family tree of that portion of the $(j - 1)$th segment contiguous to the jth segment. The key observation is that, if no recombination events have occurred in generations $1, \ldots, t - 1$ in the ith segment, then the history of the ith segment up to generation t is exactly the same as the contiguous portion of the $(i + 1)$th segment. Thus, given $l(0), \ldots, l(t - 1)$, the probability that no recombination events occur in generation t in the ith segment is $(1 - c/m)^{l(t-1)}$. Define,

$$\delta_i(t) = \begin{cases} 1 & \text{if no recombination event occurred in the} \\ & i\text{th segment in generation } t. \\ 0 & \text{otherwise.} \end{cases}$$

Let $\delta_j(t)$ be defined analogously for the jth segment. It follows that

$$\begin{aligned} P(R_i = 0) &= P\left(\prod_t \delta_i(t) = 1\right) \\ &= E\left(\prod_t P\left(\delta_i(t) = 1 \,\middle|\, \prod_{k=0}^{t-1} \delta_i(k) = 1, \underset{\rightarrow}{l}\right)\right) \\ &= E\left(\prod_t \left(1 - \frac{c}{m}\right)^{l(t)}\right) \\ &= E\left(\left(1 - \frac{c}{m}\right)^{T_i}\right) \end{aligned} \tag{8}$$

where $T_i = \sum_t l(t)$. Similarly

$$P(R_j = 0) = E\left(\left(1 - \frac{c}{m}\right)^{T_j}\right) \tag{9}$$

where $T_j = \sum_t r(t)$. Also, if no recombination events have occurred in generations 1, . . . $t - 1$ in the ith and jth segments, then, given $l(0)$, $r(0)$, . . . , $l(t-1)$, $r(t-1)$, the probability that no recombination events occurred in the ith and jth segments in generation t is $(1 - c/m)^{l(t-1)+r(t-1)}$. Thus,

$$\begin{aligned} P(R_i = 0,\ R_j = 0) &= E\left(\prod_t \delta_i(t)\delta_j(t)\right) \\ &= E\left(\prod_t E\left(\delta_i(t)\delta_j(t) \,\middle|\, \prod_{k=0}^{t-1} \delta_i(k)\delta_j(k) = 1,\ \underrightarrow{r},\ \underrightarrow{l}\right)\right) \qquad (10) \\ &= E\left(\left(1 - \frac{c}{m}\right)^{T_i+T_j}\right). \end{aligned}$$

From (8), (9) and (10), one concludes that

$$\begin{aligned} P(R_i = 1,\ R_j = 1) &= E\left(\left(1 - \left(1 - \frac{c}{m}\right)^{T_i}\right)\left(1 - \left(1 - \frac{c}{m}\right)^{T_j}\right)\right) \\ &\qquad - P(R_i \geq 1,\ R_j \geq 1;\ R_i + R_j \geq 3) \\ &= \left(\frac{C}{m}\right)^2 E(T_i' T_j') - P(R_i \geq 1,\ R_j \geq 1;\ R_i + R_j \geq 3) + 0(m^{-3}) \end{aligned}$$

where T_i' and T_j' are measured in units of $4N$. In APPENDIX 1 it is shown that

$$E(R_i R_j) = P(R_i = 1,\ R_j = 1) + 0(m^{-3})$$

and

$$P(R_i + R_j \geq 3) \leq 0\left(\left(\frac{2C}{m}\right)^3\right).$$

Hence,

$$E(R_i R_j) = \left(\frac{C}{m}\right)^2 E(T_i' T_j') + 0(m^{-3});$$

therefore,

$$\operatorname{cov}(R_i,\ R_j) = \left(\frac{C}{m}\right)^2 \operatorname{cov}(T_i', T_j') + 0(m^{-3}). \qquad (11)$$

T_i' and T_j' are quantities related to family trees of two parts of the genome for which the recombination rate between them is $((j - i)/m)C$. Thus, $\operatorname{cov}(T_i', T_j')$ can be studied using the two-locus theory. In particular, HUDSON (1983b) has shown that $\operatorname{cov}(T_i', T_j') = f_n(((j - i)/m)C)$, where f_n is a function that depends on the sample size.

We can now evaluate $\operatorname{Var}(R)$. From (6), (7), (10) and (11) we obtain

$$\operatorname{Var}(R) = C\left(\sum_{i=1}^{n-1} \frac{1}{i}\right) + 2\left(\frac{C}{m}\right)^2 \sum_{i=1}^{m} (m - i)\, f_n\left(\frac{i}{m}\, C\right) + 0(m^{-1}). \qquad (12)$$

Letting $m \longrightarrow \infty$ results in the formula:

$$\mathrm{Var}(R) = C\left(\sum_{i=1}^{n-1} \frac{1}{i}\right) + 2\int_0^C (C - z)f_n(z)dz. \tag{13}$$

For $n = 2$, HUDSON has shown that

$$f_2(z) = \frac{z + 18}{z^2 + 13z + 18}.$$

For $n \geq 3$ no formula is known for $f_n(z)$, but N. L. KAPLAN and R. R. HUDSON (unpublished results) have recently shown how to compute $f_n(z)$ for any values of n and z. They also proposed the following two approximations for $f_n(z)$:

$$f_n(z) \approx \left(\sum_{i=1}^{n-1} \frac{1}{i^2}\right) f_2(z) \tag{14}$$

and

$$f_n(z) \approx \frac{n}{2z(n-1)}. \tag{15}$$

The estimate in (14) is more appropriate for small to moderate values of z, whereas the estimate in (15) is more accurate for larger values of z (N. L. KAPLAN and R. R. HUDSON, unpublished results).

It is interesting to note the similarity between the formulas for the mean and variance of R and S, the number of segregating sites in the sample. WATTERSON (1975) has shown that

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i},$$

and HUDSON (1983b) has proved that

$$\mathrm{Var}(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \frac{2\theta^2}{C^2}\int_0^C (C - z)f_n(z)dz.$$

Thus, when $\theta = C$, $E(S) = E(R)$ and $\mathrm{Var}(S) = \mathrm{Var}(R)$. Even though the mean and variance of S and R are equal when $\theta = C$, simulation results not shown indicate that the distributions of S and R are not the same.

STATISTICAL PROPERTIES OF R_M

Before the statistical properties of R_M are explored, it is instructive to elaborate on its definition. It is clear from the discussion of the four-gamete test that the sample size must be at least four and only segregating sites in the sample need to be compared. Suppose there are S segregating sites in the sample labeled 1 to S. Define,

$$d(i, j) = \begin{cases} 1 & \text{if all four gametes are present in the} \\ & \text{sample for sites } i \text{ and } j. \\ 0 & \text{otherwise.} \end{cases}$$

and let $\mathbf{D} = (d(i, j))_{1 \leq i, j \leq S}$. The value of R_M represents the minimum number

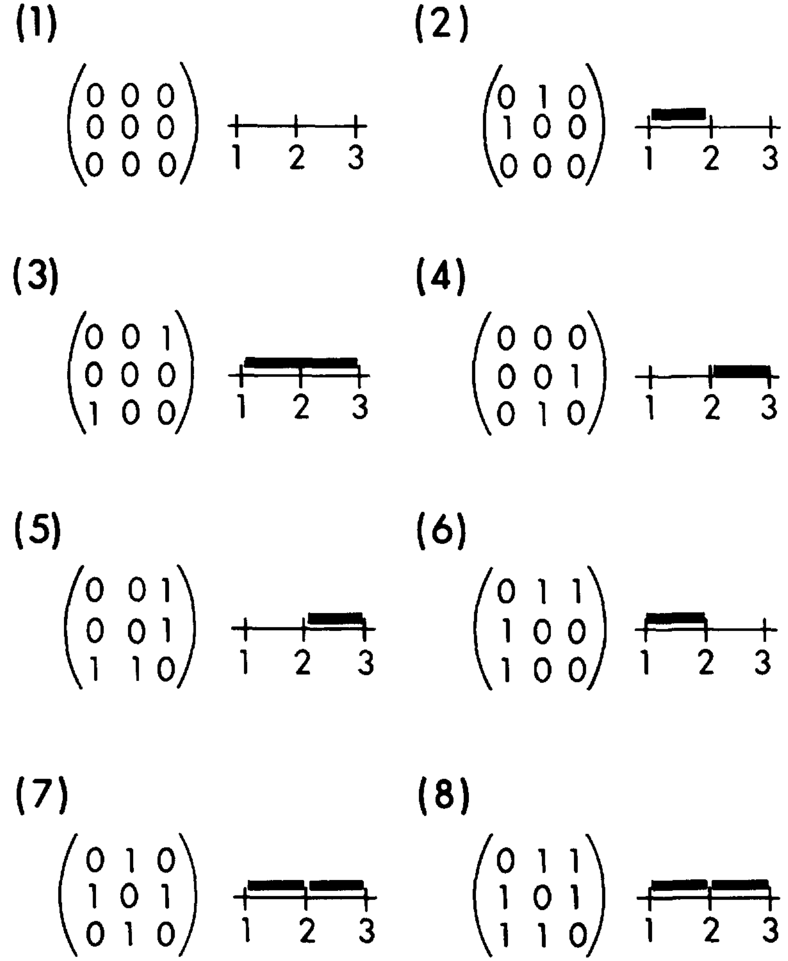

FIGURE 1.—The possible **D** matrices when there are three segregating sites in the sample. To the right of each matrix is a representation of the chromosome indicating with a solid bar the regions in which at least one recombination event must have occurred to account for the pairs of sites that pass the four-gamete test.

of recombination events in the history of the sample which is consistent with the structure of **D**. To demonstrate this idea we consider the following simple example. Suppose there are three segregating sites in the sample, then there are three pairs of sites to which the four-gamete test can be applied. In Figure 1 the spatial relationship between the recombination events and the segregating sites is given for the eight possible **D** matrices. In case (1) there is no evidence of recombination and, therefore, $R_M = 0$. For cases (2) through (6) only one event is needed to account for the pairs of sites that pass the four-gamete test; for cases (7) and (8) two recombination events are required. For large numbers of segregating sites the determination of R_M from **D** is not a simple matter and an algorithm for doing this is given in APPENDIX 2.

It is clear from the definition of R_M that increasing S cannot decrease R_M.

Thus, $E(R_M \mid S)$ is a nondecreasing function of S. Furthermore, HUDSON (1983b) has shown that distribution of S conditional on the history of the sample is Poisson with mean $\theta T/4N$, where T is a quantity that depends only on the history of the sample. Thus,

$$E(R_M) = E(E(R_M \mid S)) = E\left(\sum_j e^{-\theta T/4N}\left(\frac{\theta T}{4N}\right)^j E(R_M \mid j)\right).$$

Since $E(R_M \mid j)$ is a nondecreasing function of j, $E(R_M)$ is an increasing function of θ (BARLOW and PROSCHAN 1975). Define

$$F_n(\theta) = \frac{E(R_M)}{E(R)}.$$

One is tempted to interpret $F_n(\theta)$ as the probability that a particular recombination event is detected by the four-gamete test when the mutation rate is θ. This interpretation can be justified when θ is large. Consider a segment of length $1/m$ and let R' and R'_M denote the values of R and R_M for this segment. It is shown in APPENDIX 3 that

$$F_n(\infty) = \lim_{\theta \to \infty} F_n(\theta) = \lim_{m \to \infty} \lim_{\theta \to \infty} P(R'_M = 1 \mid R' = 1). \tag{16}$$

The existence of $F_n(\infty)$ is immediate since $E(R)$ does not depend on θ, and $E(R_M)$ is an increasing function of θ. In APPENDIX 3 it is shown that

$$F_n(\infty) = \frac{16}{\sum_{i=1}^{n-1} \frac{1}{i}}\left[\sum_{k=4}^{n} \frac{1}{(k+1)k^2(k-1)^2}\left[\sum_{i=2}^{k-2} \frac{1}{i}\left[\sum_{j=2}^{i} j^2(j+1)\right]\right]\right]. \tag{17}$$

As n increases, $F_n(\infty)$ converges to 1 and it appears from calculations that it does so in a monotonic fashion. The rate of convergence of $F_n(\infty)$ to 1 is very slow. For example, $F_{500}(\infty) = 0.69$ and $F_{1000}(\infty) = 0.71$. There is no obvious way to compute $F_n(\theta)$ for finite values of θ and, therefore, simulation methods must be used to estimate it.

SIMULATION RESULTS

Simulations were carried out to study the mean and variance of R and R_M. The history of the sample was generated using HUDSON'S (1983b) algorithm. Given the collection of family trees, the number of mutations on any branch has a Poisson distribution with mean θt_i, where t_i is the length of the branch. Furthermore, each mutation is independently and uniformly placed on the branch. In Tables 1, 2 and 3 the estimated means and variances of R and R_M obtained with Monte Carlo simulations are shown for sample sizes of 11 and 25. The sample size of 11 was chosen to match the sample size in the data set of KREITMAN (1983) which is discussed in the next section.

The estimated mean of R is consistent with the predicted mean given by (5). The estimated variance of R is in good agreement with the predicted values obtained by N. L. KAPLAN and R. R. HUDSON (unpublished results).

TABLE 1

The mean and variance of R, *the number of recombination events in the history of the sample*

Sample size	Statistic	C 1	5	10	20	50	100
11	$\hat{E}(R)^a$	2.94	14.5	29.2	58.6	146.6	292
11	$E(R)^b$	2.93	14.6	29.3	58.6	146.5	293
11	$\widehat{\mathrm{Var}}(R)^c$	4.29	33.5	86.7	226	655	1497
11	$\mathrm{Var}(R)^d$	4.21	34.9	87.7	215	665	1509
11	$\widetilde{\mathrm{Var}}(R)^e$	4.21	35.4	89.7	223	708	1645
25	$\hat{E}(R)^a$	3.84	18.7	37.5	74.7	189	378
25	$E(R)^b$	3.78	18.9	37.8	75.5	189	378
25	$\widehat{\mathrm{Var}}(R)^c$	5.10	39.2	99.5	255	795	1683
25	$\mathrm{Var}(R)^d$	5.10	40.1	99.4	242	752	1715
25	$\widetilde{\mathrm{Var}}(R)^e$	5.11	40.3	100.3	245.3	770	1777

[a] An estimate of $E(R)$ obtained from 1000 computer-generated samples.

[b] $E(R) = C \sum_{j=1}^{n-1} \frac{1}{j}$.

[c] An estimate of Var(R) obtained from 1000 computer-generated samples.

[d] Calculated using the results of N. L. KAPLAN and R. R. HUDSON (unpublished results).

[e] $\widetilde{\mathrm{Var}}(R) = C \sum_{j=1}^{n-1} \frac{1}{j} + 2\left(\sum_{j=1}^{n-1} \frac{1}{j^2}\right) \int_0^C (C - z) f_2(z)dz$ which is obtained from (13) and (14).

TABLE 2

Monte Carlo estimates of the mean and variance of R_M *in samples of size 11*

θ	Statistic	C 1	5	10	20	50	100
5	$\hat{E}(R_M)^a$	0.17	0.69	1.1	1.6	2.6	3.3
5	$\widehat{\mathrm{Var}}(R_M)^a$	0.17	0:62	0.85	1.2	2.0	2.4
10	$\hat{E}(R_M)^a$	0.30	1.1	1.8	2.8	4.5	6.0
10	$\widehat{\mathrm{Var}}(R_M)^a$	0.31	0.99	1.4	1.9	3.0	4.5
15	$\hat{E}(R_M)^a$	0.39	1.4	2.3	3.6	5.9	7.9
15	$\widehat{\mathrm{Var}}(R_M)^a$	0.37	1.1	1.9	2.7	4.1	5.5
30	$\hat{E}(R_M)^a$	0.51	2.0	3.3	5.1	8.6	12
30	$\widehat{\mathrm{Var}}(R_M)^a$	0.46	1.8	3.0	4.2	6.8	9.7
[c]	$\hat{E}(R_M)^b$	0.42	1.5	2.3	3.6	5.8	8.0
[c]	$\widehat{\mathrm{Var}}(R_M)^b$	0.40	1.1	1.4	2.0	2.9	3.6

[a] Estimates of the mean and variance of R_M obtained from 1000 computer-generated samples.

[b] Estimates of the mean and variance of R_M conditional on 44 segregating sites in the sample obtained from 1000 computer-generated samples.

[c] Number of segregating sites was fixed at 44.

TABLE 3

Monte Carlo estimates of the mean and variance of R_M *in samples of size 25*

		C					
θ	Statistic	1	5	10	20	50	100
5	$\hat{E}(R_M)^a$	0.37	1.2	1.9	2.7	4.0	5.3
5	$\widehat{\text{Var}}(R_M)^a$	0.33	0.92	1.4	1.9	2.5	3.4
10	$\hat{E}(R_M)^a$	0.54	1.9	2.9	4.4	6.7	8.9
10	$\widehat{\text{Var}}(R_M)^a$	0.46	1.3	2.0	2.8	4.2	4.9
15	$\hat{E}(R_M)^a$	0.67	2.4	3.7	5.5	8.8	11
15	$\widehat{\text{Var}}(R_M)^a$	0.57	1.7	2.4	3.5	5.5	7.0
30	$\hat{E}(R_M)^a$	0.97	3.3	5.2	8.1	13	18
30	$\widehat{\text{Var}}(R_M)^a$	0.84	2.5	4.3	5.8	8.9	12
[c]	$\hat{E}(R_M)^b$	0.61	2.2	3.2	4.8	7.4	9.8
[c]	$\widehat{\text{Var}}(R_M)^b$	0.47	1.2	1.6	2.3	3.3	3.8

[a] Estimates of the mean and variance of R_M obtained from 1000 computer-generated samples.

[b] Estimates of the mean and variance of R_M conditional on 44 segregating sites in the sample obtained from 1000 computer-generated samples.

[c] Number of segregating sites was fixed at 44.

The simulation results in Tables 1, 2 and 3 show that the expectation of R_M is much smaller than the expectation of R. Even with C small and θ large, the case in which R_M will be closest to R, the mean of R_M is much smaller than the mean of R. For example, when $n = 11$, $C = 1$ and $\theta = 30$, the mean of R_M was 0.506 and the mean of R was 2.94. Thus, the probability that any particular recombination event is detected using the four-gamete test is only about $0.506/2.94 = 0.17$. It follows from (16) that as θ approaches infinity, this probability will approach 0.32. For smaller mutation rates and higher recombination rates, the ratio of the mean of R_M to the mean of R is even smaller. For example, with $n = 11$, $C = 20$ and $\theta = 15$, the ratio is $3.62/58.3 = 0.06$. The variance of R_M for which no formula is known is approximately equal to the mean of R_M for $C < 10$, and for $C > 10$, is somewhat smaller than the mean.

The simulations just described were carried out with a specified value of θ, and, therefore, the number of segregating sites in each sample is a random quantity. Since the number of segregating sites in a sample is observable and θ is unknown, it may be appropriate for the interpretation of data to consider the distribution of R_M for a given number of segregating sites in the sample. This conditional distribution is not difficult to study since, given the collection of family trees, each of the given numbers of mutations is independently and uniformly placed on a branch of length t_i with probability $t_i/\sum_i t_i$. Shown near the bottom of Tables 2 and 3 are the means and variances of R_M for several rates of recombination and 44 segregating sites. This number of segregating sites was chosen to match the number of segregating sites in the data of

TABLE 4

Estimates of the probability of detecting a single recombination event in samples of size 11 and 25 when specified numbers of sites are segregating on each side of the recombination site

No. of segregating sites on each side of the recombination site	Probability of detection[a]	
	$n = 11$	$n = 25$
10	0.12	0.16
20	0.22	0.26
50	0.29	0.36
100	0.31	0.43
200	0.32	0.43
∞	0.318[b]	0.450[b]

[a] Each estimate was obtained from 1000 computer-generated samples.

[b] Calculated with (16).

KREITMAN (1983). The variance of R_M tends to be somewhat reduced when the number of segregating sites is fixed. This is not unexpected since one component of the variability of R_M is removed if the number of segregating sites is fixed.

Table 4 gives Monte Carlo estimates of the probability that a single recombination event is detected using the four-gamete test when given numbers of mutations occur on each side of the recombination site. With ten mutations on each side of the recombination site, less than half of those recombination events that are potentially detectable are detected. With 100 mutations on each side, the asymptotic value, as predicted by (16), is nearly reached.

KREITMAN DATA

KREITMAN (1983) published the DNA sequences of the *Adh* region from 11 chromosomes obtained from natural populations of *D. melanogaster*. Although the 11 chromosomes were not a random sample, in this section we will treat them as if they were. There are 43 polymorphic nucleotide sites in the sample. For these 43 sites, $R_M = 4$. In addition, there are four sites with insertion/deletion polymorphisms, each of which can be interpreted as resulting from a single insertion/deletion event. If these insertion/deletion polymorphisms are included, then there are 47 polymorphic sites, and $R_M = 5$. It follows from WATTERSON'S (1975) results that an estimate of θ is $S/(\sum_{j=1}^{10} 1/j)$, where S is the number of segregating sites in the sample. For Kreitman's data $S = 43$ or 47, depending on whether insertion/deletion polymorphisms are included and, therefore, an estimate of θ is approximately 15. Estimates of θ from electrophoretic data are much smaller, typically about 0.1. This is not surprising since electrophoresis will not detect variation in untranslated regions, at silent sites and often even at nonsilent sites. If, however, some of the segregating sites are caused by deleterious mutations, then the above estimate of θ may be biased upward. Table 2 shows that, if $\theta = 15$, then the value of C for which

$E(R_M) = 4$ is less than 50 and greater than 20. One cannot conclude with any degree of confidence that C lies between these two values because the variance of R_M is large over this range of C values. However, it can be argued that C is likely to be between 5 and 150. Indeed, with $C = 5$ and the number of segregating sites fixed at 43, only eight of 1000 stimulated samples had $R_M \geq 4$, and with the number of segregating sites fixed at 47, only 12 of the 1000 samples had $R_M \geq 5$. On the other hand, with $C = 150$ and 43 segregating sites, R_M was ≤ 4 for only ten of 1000 samples, and with 47 segregating sites, R_M was ≤ 5 for 15 of 1000 samples.

Is an estimate of C between 5 and 150 compatible with other estimates of N and c? The average recombination rate per kilobase in *D. melanogaster* females has been estimated to be 1.7×10^{-5} (CHOVNICK, GELBART and MCCARRON 1977). Since there is essentially no recombination in *Drosophila* males and the sequence data covers 2.7 kb, an estimate of c is (1.7×10^{-5}) (0.5) $(2.7) = 2.3 \times 10^{-5}$. If C lies between 5 and 150, then an estimate of N is between 5.4×10^4 and 1.6×10^6. This range of values is well below Kreitman's estimate, 3.3×10^6, which was based on the number of segregating sites and estimates of the mutation rate. This discrepancy may be due just to the sampling variability in the number of segregating sites or the estimate of the neutral mutation rate used by Kreitman may be too small or the recombination rate in the *Adh* region may be smaller than 1.7×10^{-5}/kb.

DISCUSSION

In some ways, R, the number of recombination events in the history of a sample, and S, the number of segregating sites in a sample, are analogous. One can intepret S as the number of mutation events in the history of the sample. The similar formulas for the means and for the variances of R and S reflect how analogous the two quantities are. It is well known that $4Nu$ can be effectively estimated using S (WATTERSON 1975; EWENS 1979). Unfortunately, $4Nc$ cannot be estimated in an analogous fashion using R, because R is not directly observable in samples. An observable quantity that is related to R is R_M. However, R_M is typically much smaller than R and has statistical properties that make estimates of $4Nc$ imprecise. The mean of R_M grows rather slowly with increasing values of $4Nc$ and the variance is large enough so that a large range of $4Nc$ values are compatible with an observed value of R_M. Nevertheless, some information about $4Nc$ can be obtained using R_M as was illustrated for the KREITMAN (1983) data. An estimate of N, the population size, can be obtained if c is known. Direct estimates of c can be obtained for some genetic loci (CHOVNICK, GELBART and MCCARRON 1977). Such estimates of c are likely to be much more precise than estimates of u, which are typically based on estimated divergence times of species. Consequently, estimates of N based on R_M may still be more precise than estimates based on S. It should be noted, however, that such estimates may be quite sensitive to departures from random mating and constancy of population size.

Our results also bear on the problem of reconstructing phylogenies of DNA sequences. For most samples, R_M is much smaller than R. This means that for

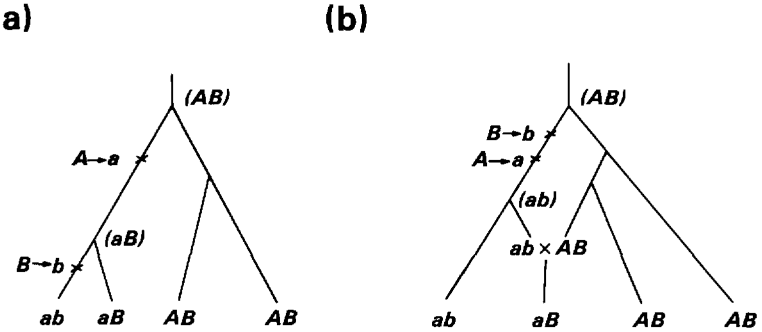

FIGURE 2.—Two possible histories of a sample of four gametes. (a) In this history no recombination events occur. This is a most parsimonious tree relating the four gametes. (b) In this history one recombination event occurs. Although the resulting gametes are the same, the order of the mutations, the placement of the mutations on the tree and the ancestral genotypes are different from the history in (a).

a typical sample many more recombination events probably took place in the descent from the most recent common ancestors of the sample than would appear in a most parsimonious reconstruction of the history of the sample. Thus, when recombination can occur, it appears that parsimonious reconstructions of phylogenies of DNA segments should be viewed with some scepticism.

For some questions, undetected recombination events may be of no importance, and for others they may be relevant. For example, if one is concerned with the order in which mutations arose, or with the actual genotypes of the ancestors of sampled gametes, then conclusions based on parsimonious reconstructions can be incorrect. An example in which four gametes are sampled is shown in Figure 2. The four gametes are denoted *aB*, *ab*, *AB* and *AB*. Suppose that it is known that the ancestral state is *AB* (from the examination of related species, for example). As shown in Figure 2a, no recombination events are required to explain the origin of the four gametes. If the ancestral state is *AB*, the most parsimonious tree, without recombination, leads to the conclusion that the mutation $A \rightarrow a$ occurred before the mutation $B \rightarrow b$ and that the genotype *Ab* never existed in the ancestry of the sample. However, Figure 2b shows that, with a recombination event, the order and placement of the mutations could be much different, with the mutation $B \rightarrow b$ occurring before the mutation $A \rightarrow a$ and the genotype *Ab* being an ancestor of the *ab* and the *aB* gametes.

Our simulations show that even with fairly large numbers of mutations many recombination events will go undetected. For example, with 44 segregating sites and $C = 40$, in samples of size 11, the mean value of R_M is approximately 5, whereas the mean number of recombination events is 118. These parameter values were chosen to match the data of KREITMAN (1983). Reconstructions of the history of Kreitman's sample of gametes that contain only five recombination events are, therefore, likely to be quite misleading.

LITERATURE CITED

BARLOW, R. E. and F. PROSCHAN, 1975 *Statistical Theory of Reliability and Life Testing Probability Models.* Holt, Rinehart and Winston Inc., New York.

CHOVNICK, A., W. GELBART and M. MCCARRON, 1977 Organization of the Rosy locus in *Drosophila melanogaster*. Cell **11:** 1–10.

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

GRIFFITHS, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Pop. Biol. **19:** 169–186.

HUDSON, R. R., 1983a Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1983b Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23:** 183–201.

KIMURA, M., 1971 Theoretical foundations of population genetics at the molecular level. Theor. Pop. Biol. **2:** 174–208.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

TAJIMA, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics **105:** 437–460.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256–276.

Communicating editor: B. S. WEIR

APPENDIX 1

To study properties of the distribution of R_1 it is necessary to describe in more detail the method used by HUDSON (1938b) to simulate the history of a sample. We first introduce two definitions. If any two ancestral gametes of segment 1 in generation $t - 1$ have a common ancestor in generation t, we say that a CA event has occurred in generation t. Alternatively, if an ancestral gamete of segment 1 in generation $t - 1$ is the recombinant descendent of two ancestral gametes of segment 1 in generation t, we say that an RE event has occurred in generation t. Let event i be the ith most recent event. HUDSON (1983b) has shown that the distribution of the time (measured in units of $4N$) between the $(i - 1)$th and ith events $(i \geq 1)$ conditioned on the history of the sample for segment 1 up to the time of the $(i - 1)$th event is asymptotically exponential with parameter $K_i(K_i - 1) + G_iC/m$, where K_i represents the number of ancestral gametes of segment 1 in the generation when the ith event occurred (note that $K_i \geq 2$) and G_i is a number that lies between 0 and K_i and depends on the outcomes of the previous $i - 1$ events. The quantity G_ic/m is the probability that an RE event occurs in any particular generation between the $(i - 1)$th event and the ith event. Furthermore, the probability that the ith event is a CA event is $K_i(K_i - 1)/(K_i(K_i - 1) + G_iC/m)$, and the probability that it is an RE event is $(G_iC/m)/(K_i(K_i - 1) + G_iC/m)$.

Define:

$$\eta_i = \begin{cases} 1 & \text{if } i\text{th event is an RE event.} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$P(R_1 = j) = P\left(\sum_i \eta_i = j\right).$$

The only value of j for which $P(R_1 = j)$ can be evaluated is $j = 0$. In this case all the events are CA events, and

$$\frac{G_iC}{m} = (n - i + 1)\frac{C}{m} \text{ and } K_i = n - i + 1, \qquad 1 \le i \le n - 1.$$

Thus,

$$\begin{aligned} P(R_1 = 0) &= P\left(\sum_{i=1}^{n-1} \eta_i = 0\right) \\ &= P(\eta_1 = 0) \prod_{i=2}^{n-1} P(\eta_i = 0 \mid \eta_1 = \cdots = \eta_{i-1} = 0) \\ &= \prod_{i=2}^{n} \frac{i(i-1)}{i\frac{C}{m} + i(i-1)} \\ &= \prod_{i=2}^{n} \frac{i-1}{\frac{C}{m} + i - 1} \ . \end{aligned}$$

For values of $j \ge 1$ no simple formula for $P(R_1 = j)$ exists. However, we do have the following inequality.

Lemma. $P(R_1 = j) \le (C/m)^j$; $j \ge 2$.

Proof. For simplicity we assume that $j = 2$. The argument for larger values of j is the same. Let $1 \le j_1 < j_2$. Then,

$P(\eta_{j_1} = 1,\ \eta_{j_2} = 1$ and all the other $\eta_i = 0)$

$$= \left(\frac{(G_{j_1}C)/m}{(G_{j_1}C)/m + K_{j_1}(K_{j_1} - 1)}\right)\left(\frac{(G_{j_2}C)/m}{(G_{j_2}C)/m + K_{j_2}(K_{j_2} - 1)}\right) \prod\nolimits_{i \ne j_1 j_2} \left(\frac{K_i(K_i - 1)}{(G_iC)/m + K_i(K_i - 1)}\right) \le \left(\frac{C}{m}\right)^2$$

$P($all η_i are zero, $i \ne j_1, j_2$, and no conditions are placed on η_{j_1} and $\eta_{j_2})$.
The last inequality follows because j_1 and $j_2 \ge 2$, $G_{j_1} \le K_{j_1}$, and $G_{j_2} \le K_{j_2}$. Summing over all j_1 and j_2 results in

$$P(R_1 = 2) \le \left(\frac{C}{m}\right)^2 P(R_1 \le 2) \le \left(\frac{C}{m}\right)^2.$$

It follows from the lemma that

$$\begin{aligned} E(R_1, R_1 \ge 2) &= \sum_{j \ge 2} jP(R_1 = j) \\ &\le \left(\frac{C}{m}\right)^2 \sum_{j \ge 2} j \left(\frac{C}{m}\right)^{j-2} \\ &\le \frac{2\left(\frac{C}{m}\right)^2}{1 - \frac{C}{m}} = 0\left(\left(\frac{C}{m}\right)^2\right). \end{aligned}$$

The argument for showing that $E(R_iR_j,\ R_i + R_j \ge 3) = 0((C/m)^3)$ is similar to the one above. For this case one considers the history of the entire segment from the ith segment to the jth segment. The same structure governs the interevent times except that now one needs to also keep track of where along the segment the RE events occur. When an RE event occurs in generation t the ancestral gamete in generation $t - 1$ is divided into two pieces at a randomly chosen crossover point. Hence, $P[(l + 1)$th event is an RE event and it occurred in either the ith or jth segment | the history of the sample up to the lth event$] \le (2G_lC^2)/[mG_lC + m^2K_l\,(K_l - 1)]$.

The argument used in the lemma can be used to show that

$$P(R_i + R_j = l) \le \left(\frac{2C}{m}\right)^l; \qquad l \ge 3.$$

Thus,

$$\begin{aligned} E(R_iR_j;\ R_i + R_j \ge 3) &\le \sum_{k+l\ge 3}\sum kl\left(\frac{2C}{m}\right)^{k+l} \\ &\le \left(\frac{2C}{m}\right)^3 \sum_{j\ge 3} j^3 \left(\frac{2C}{m}\right)^{j-3} \\ &= 0\left(\left(\frac{2C}{m}\right)^3\right). \end{aligned}$$

APPENDIX 2

We describe here an algorithm for determining R_M. Recall the matrix $\mathbf{D} = (d(i, j))$, where $d(i, j) = 1$ if all four gametes involving sites i and j are present in the sample; otherwise $d(i, j) = 0$. To each nonzero element $d(i, j)$ of $\mathbf{D}$ above the diagonal we associate the open interval (i, j) and form a list of these intervals ordered so that the starting points of the intervals are not decreasing. The method for finding R_M deletes certain members of this list. The first type of intervals deleted are those that completely contain other intervals. For example, if (i, j) and (m, n) are on the list, and $m \le i < j \le n$, then (m, n) is deleted. For the remaining intervals on the list, let (i_1, j_1) be the first interval that is not disjoint from all of the others. All intervals (m, n) such that $i_1 < m < j_1$ are then deleted. Next, let (i_2, j_2) be the first interval with $i_2 \ge j_1$ such that (i_2, j_2) is not disjoint from all of the remaining intervals and delete all intervals whose first component is $<j_2$ and $>i_2$. This process is continued until it is not possible to find an interval that has a nonempty intersection with some other interval on the list. At this point all of the intervals are disjoint, and to each interval we must assign at least one recombination event. Hence, R_M equals the final number of intervals on the list.

APPENDIX 3

Let

$$\Lambda = 16 \sum_{k=4}^{n} \frac{1}{(k + 1)k^2(k - 1)^2}\left[\sum_{i=2}^{k-2} \frac{1}{i}\left[\sum_{j=2}^{i} j^2\,(j + 1)\right]\right]$$

Since $E(R) = C(\sum_{i=1}^{n-1} 1/i)$, it is sufficient to establish (17) to show that

$$\lim_{\theta\to\infty} E(R_M) = C\Lambda.$$

If we divide the genome up into m segments, then

$$\lim_{\theta\to\infty} E(R_M) = \sum_{i=1}^{m} E(R_M(i)) = mE(R_M(1)),$$

where $R_M(1)$ represents the number of recombination events in segment 1 in the history of the sample which are potentially detected by the four-gamete test, *i.e.*, they are detectable when the number of segregating sites is sufficiently large. If $R(1)$ equals the number of recombination events in segment 1 in the history of the sample, then $R_M(1) \le R(1)$ and, therefore,

$$E(R_M(1)) = P(R_M(1) = 1,\ R(1) = 1) + 0(m^{-2}).$$

To evaluate $P(R_M(1) = 1, R(1) = 1)$ we observe that

$$P(R_M(1) = 1,\ R(1) = 1) = \sum_{k=4}^{n} P(k)P_k$$

where

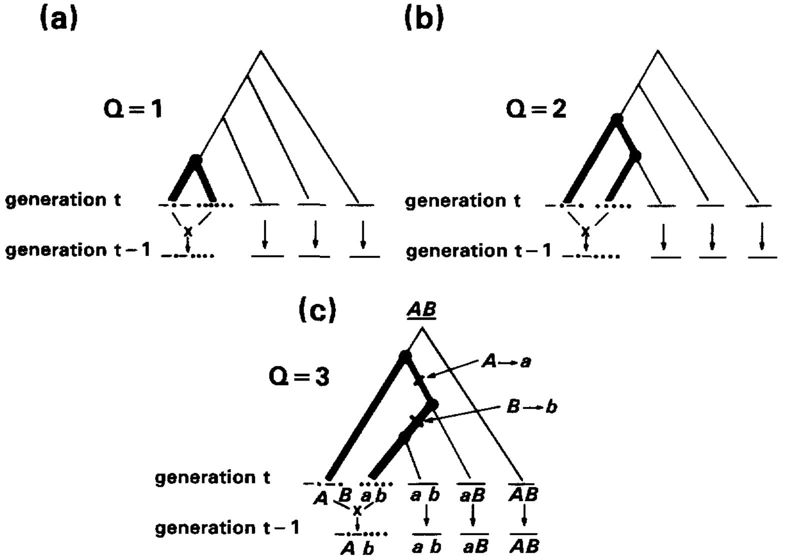

FIGURE 3.—Three possible histories of a sample of four gametes, each with one recombination event at generation t. In (a) and (b), $Q = 1$ and 2 respectively, and it is not possible to place two mutations on these trees so as to produce four gametic types in generation $t - 1$. In (c), $Q = 3$, and it is shown how two mutations can be placed so as to produce four gametic types in generation $t - 1$.

$P(k)$ = probability that the first $n - k$ events in the history of the sample for segment 1 are CA events and the $(n - k + 1)$th is the RE event,

and

P_k = probability that the recombination event is potentially detectable given that first $n - k$ events in the history of the sample of segment 1 are CA events and the $(n - k + 1)$th event is the RE event.

It follows from APPENDIX I that

$$P(k) = \left(\frac{\frac{C}{m}}{\frac{C}{m} + k - 1} \right) \left[\prod_{i=k+1}^{n} \left(\frac{i - 1}{\frac{C}{m} + i - 1} \right) \right]$$
$$= \frac{C}{m} \frac{1}{k - 1} + 0(m^{-2}).$$

To evaluate P_k it is necessary to determine which topologies of the family tree will result in a potentially detectable recombination event. Suppose that the first $(n - k)$ events in the history of the sample are CA events, the $(n - k + 1)$th is the RE event and this event occurs in generation t. Hence, in generation t there are $k - 1$ ancestral gametes that are direct ancestors of the gametes in the sample and two ancestral gametes that form a recombinant. The topology of the tree determined by the first $(n - k)$ events is not relevant to the question of detectability and, therefore, does not need to be considered. The possible topologies of the tree determined by the $k - 1$ CA events proceeding the RE event can be categorized in the following way. For any possible tree consider the path connecting the two gametes that form the recombinant. Let Q denote the number of nodes on this path. In Figure 3 examples of trees are given in which $Q = 1$, 2 and 3. For the RE event to be potentially detectable it must be possible to place two mutations on the tree, one to the left of the crossover point and one to the right, which result in four gametes in the sample. Since no RE events occur in the first $(n - k)$ events, it is necessary that the four gametes be present among the ancestral gametes in the generation in which the $(n - k)$th event occurred.

It is straightforward to check that, if $Q = 1$ or 2, then it is impossible to obtain four gametes in the sample regardless of where the two mutations are placed. Thus, P_k is identified as the probability that the family tree relating the $k - 1$ gametes and the two gametes that form the recombinant is such that $Q \geq 3$. A tree with $Q \geq 3$ can only occur in the following way:

(1) The first i events are CA events not involving either of the two gametes that form a recombinant. $(i = 0, \ldots, k - 4)$;

(2) The $(i + 1)$ event is a CA event involving only one of these two gametes;

(3) The next l CA events do not involve either of these two gametes or their ancestors. $(l = 0, \ldots, k - 4 - i)$; and

(4) The $(i + l + 2)$ event involves only one of these two gametes or their ancestors.

Let $P_k\ (i, l)$ denote the probability of (1)–(4) for $0 \leq i \leq k - 4$ and $0 \leq l \leq k - 4 - i$. Then

$$P_k(i, l) = \prod_{j=0}^{i-1} \left[\left(\frac{\binom{k-1-j}{2}}{\binom{k+1-j}{2}} \right) \right] \left(\frac{\binom{2}{1}\binom{k-1-i}{1}}{\binom{k+1-i}{2}} \right) \prod_{j=0}^{l-1} \left[\left(\frac{\binom{k-2-i-j}{2}}{\binom{k-i-j}{2}} \right) \right] \left(\frac{\binom{2}{1}\binom{k-l-i-2}{1}}{\binom{k-i-l}{2}} \right)$$

$$= 16 \frac{(k - i - l - 1)(k - i - l - 2)^2}{(k + 1)k^2(k - 1)(k - i - 2)}.$$

Thus,

$$P_k = \frac{16}{(k + 1)k^2(k - 1)} \sum_{i=0}^{k-4} \frac{1}{k - i - 2} \left[\sum_{l=0}^{k-4-i} (k - i - l - 2)^2(k - i - l - 1) \right]$$

$$= \frac{16}{(k + 1)k^2(k - 1)} \sum_{i=2}^{k-2} \frac{1}{i} \left[\sum_{j=2}^{i} j^2(j + 1) \right].$$

Finally we have

$$P(R_M(1) = 1, R(1) = 1) = \frac{16C}{m} \sum_{k=4}^{n} \frac{1}{(k + 1)k^2(k - 1)^2} \left[\sum_{i=2}^{k-2} \frac{1}{i} \left[\sum_{j=2}^{i} j^2(j + 1) \right] \right] + 0(m^{-2}).$$

Letting m tend to ∞ we obtain

$$\lim_{\theta \to \infty} E(R_M) = 16C \sum_{k=4}^{n} \frac{1}{(k + 1)k^2(k - 1)^2} \left[\sum_{i=2}^{k-2} \frac{1}{i} \left[\sum_{j=2}^{i} j^2(j + 1) \right] \right] = CA.$$

To prove equation 16 one only has to note that for large θ, $R'_M = R_M(1)$. Thus

$$\lim_{m \to \infty} \lim_{\theta \to \infty} P(R'_M = 1 \mid R' = 1) = \lim_{m \to \infty} \frac{P(R_M(1) = 1, R(1) = 1)}{P(R(1) = 1)} = F_n(\infty).$$