Statistical Tests of Neutrality of Mutations

Yun-Xin Fu[1] and Wen-Hsiung Li

Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77225

Manuscript received July 9, 1992
Revised copy accepted November 18, 1992

ABSTRACT

Mutations in the genealogy of the sequences in a random sample from a population can be classified as external and internal. External mutations are mutations that occurred in the external branches and internal mutations are mutations that occurred in the internal branches of the genealogy. Under the assumption of selective neutrality, the expected number of external mutations is equal to $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the rate of mutation per gene per generation. Interestingly, this expectation is independent of the sample size. The number of external mutations is likely to deviate from its neutral expectation when there is selection while the number of internal mutations is less affected by the presence of selection. Statistical properties of the numbers of external mutations and of internal mutations are studied and their relationships to two commonly used estimates of θ are derived. From these properties, several new statistical tests based on a random sample of DNA sequences from the population are developed for testing the hypothesis that all mutations at a locus are neutral.

An important issue in molecular population genetics is how to detect the presence of natural selection among the variants of a nucleotide sequence in a population (*e.g.*, see HUDSON, KREITMAN and AGUADE 1987; TAJIMA 1989). The pattern of polymorphism in a population is affected not only by mutation and random drift but also by selection. With the advent of rapid sequencing techniques, polymorphism data at the DNA level is expected to increase dramatically. Thus, there is a great need for a powerful test for the assumption of neutrality of mutations.

TAJIMA (1989) has proposed a method for the above purpose. He considered the number of segregating sites (K) and the average number of nucleotide differences between two sequences (Π_n) in a random sample of n sequences from a population. He noted that K is strongly affected by the existence of deleterious alleles because deleterious alleles are usually kept in low frequencies but K ignores the frequency of mutants. On the other hand, Π_n is not much affected by the existence of deleterious alleles because it considers the frequency of mutants. Therefore, if some of the sequences in the sample have selective effects, then the estimate of $\theta = 4N_e\mu$ based on K (WATTERSON 1975) will be different from the estimate based on Π_n (TAJIMA 1983); N_e is the effective population size and μ is the mutation rate per gene per generation. TAJIMA (1989) proposed to use the difference between these two estimates to detect selection among the sequences. The test statistic is

$$T = \frac{\Pi_n - K/a_n}{\sqrt{\mathrm{Var}(\Pi_n - K/a_n)}}, \tag{1}$$

where

$$a_n = \sum_{k=1}^{n-1} \frac{1}{k}. \tag{2}$$

In this paper, we propose a new approach. Consider the distribution of the mutations in the genealogy of a random sample of genes from the population. "Old" mutations will tend to be found in the older part of the genealogy while "new" mutations will likely be found in the younger part of the genealogy. The older part of the genealogy consists mainly of internal branches, while the younger part mainly of external branches. In the presence of purifying or negative selection there will tend to be excess of mutations in the external branches because deleterious alleles are present in low frequencies. Also there is likely to be excess of mutations in the external branches if an advantageous allele has recently become fixed in the population, because then the majority of the mutations in the population are expected to be young. On the other hand, if balancing (overdominant) selection is operating at the locus, then some alleles may be old and so there may be deficiency of mutations in the external branches. Therefore, comparing the numbers of mutations in internal and external branches with their expectations under selective neutrality should be a powerful way to detect selection. This is the idea behind the proposed tests in the paper.

[1] To whom correspondence should be sent.

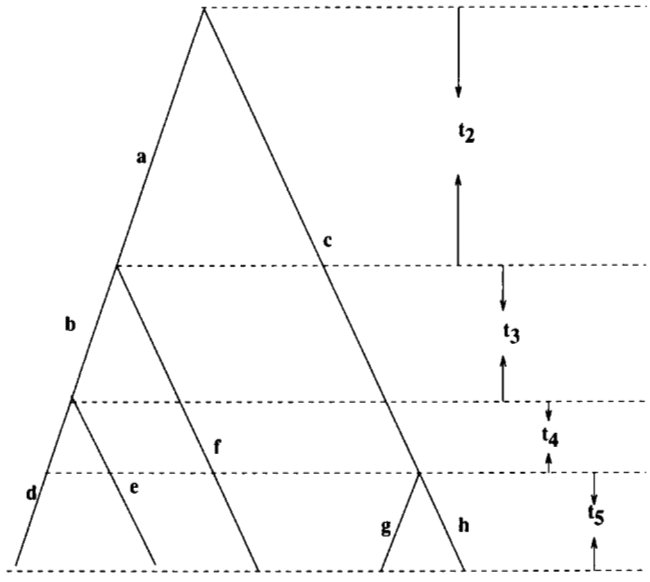


FIGURE 1.—An example of genealogy of five genes. $t_m (m = 2,\ldots,5)$ is the time (number of generations) required for the coalescence from m sequences to $m - 1$ sequences. The dashed lines signify the partition of the genealogy into $m - 1$ parts by the m branching nodes.

STATISTICAL PROPERTIES OF INTERNAL AND EXTERNAL BRANCHES

Expected time lengths of internal and external branches: Consider a random sample of n sequences of a DNA region from a diploid random mating population of effective size N_e. Assume that all mutations in the region are selectively neutral. Further, assume that the DNA region is completely linked so that no recombination occurs between sequences. Then the n sequences in the sample are connected by a single phylogenetic tree, *i.e.*, a genealogy (Figure 1). In other words, the n sequences can be traced back first to $n - 1$ ancestral sequences, next to $n - 2$ ancestral sequences and so on until reaching a single common ancestral sequence. Let t_m be the time duration (the number of generations) required for the coalescence from m sequences to $m - 1$ sequences. For convenience, we define $t_1 = 0$. KINGMAN (1982), HUDSON (1982) and TAJIMA (1983) showed that for $m \geq 2$, t_m has the exponential distribution

$$g(t_m) = \frac{m(m-1)}{4N_e} \exp\left(-\frac{m(m-1)}{4N_e} t_m\right).$$

Therefore, the first and second moments of the coalescence time t_m is

$$E(t_m) = \frac{M}{m(m-1)}, \tag{3}$$

$$E(t_m^2) = \frac{2M^2}{m^2(m-1)^2} = 2E^2(t_m), \tag{4}$$

where $M = 4N_e$.

The genealogy of n genes has $2(n - 1)$ branches. A branch is said to be external if it directly connects to an external node, otherwise it is said to be internal. Thus n of the $2(n - 1)$ branches are *external* and the other $n - 2$ branches are *internal*. Number the n external branches arbitrarily from 1 to n. The numbering is entirely for operational convenience and the number assigned to a branch does not contain any information about the relative location or the length of the branch. Let J_n, I_n and L_n be, respectively, the total time length of all branches, the total time length of internal branches and the total time length of external branches. Note that

$$J_n = I_n + L_n.$$

Let the length of the ith external branch be $l_i^{(n)}$. Then $L_n = l_1^{(n)} + l_2^{(n)} + \ldots + l_n^{(n)}$. Let l_n be the length of a randomly chosen external branch of the genealogy of n genes. Then we have

$$\begin{aligned} E(L_n) &= E(l_1^{(n)}) + \ldots + E(l_n^{(n)}) \\ &= nE(l_n), \end{aligned} \tag{5}$$

from the fact that every external branch has the same distribution because their labeling does not contain any information except for operational convenience.

The genealogy of a random sample of n genes from a single random mating population is generated by adding two external branches of length t_n to the end of a randomly chosen external branch of the genealogy $(n - 1)$ genes, while the remaining $(n - 2)$ external branches each grow by a length t_n. Therefore, we have the recurrent relationship

$$l_n = \begin{cases} l_{n-1} + t_n, & \Pr = \dfrac{n-2}{n} \\ t_n. & \Pr = \dfrac{2}{n}. \end{cases} \tag{6}$$

It follows that

$$\begin{aligned} E(l_n) &= \frac{n-2}{n}[E(l_{n-1} + t_n)] + \frac{2}{n}E(t_n) \\ &= \frac{n-2}{n}E(l_{n-1}) + E(t_n). \end{aligned} \tag{7}$$

Let $g_n = n(n - 1)l_n$. Then from Equations 1 and 5 we have

$$\begin{aligned} E(g_n) &= n(n-1)E(t_n) + E(g_{n-1}) \\ &= M + E(g_{n-1}) \\ &= M(n-1). \end{aligned} \tag{8}$$

Therefore,

$$E(l_n) = \frac{E(g_n)}{n(n-1)} = \frac{M}{n}, \tag{9}$$

$$E(L_n) = nE(l_n) = M. \tag{10}$$

Note that (10) is independent of the sample size n. That is, regardless of the number of sequences sampled, the expected total time length of the external branches is always $4N_e$ generations.

The recurrent relationship for J_n is simple. By adding one gene to the genealogy of $n-1$ genes, the total length of the genealogy increase by nt_n. From Equation 1, we have

$$\begin{aligned} E(J_n) &= E(J_{n-1}) + \frac{M}{n-1} \\ &= Ma_n. \end{aligned} \tag{11}$$

Therefore,

$$E(I_n) = E(J_n) - E(L_n) = M(a_n - 1). \tag{12}$$

Variances and covariance of internal and external branches: We have demonstrated the power of recurrent relationships for derivation of the expectations of the total time lengths of external and internal branches. The method can also be used to derive higher moments of these quantities. The variance of J_n can be directly computed. Since

$$\begin{aligned} E(J_n^2) &= E\left[\left(\sum_k kt_k\right)^2\right] \\ &= \sum_{i\neq j} ijE(t_it_j) + \sum_k k^2E(t_k^2) \\ &= \left(\sum_k kE(t_k)\right)^2 + \sum_k k^2E^2(t_k) \\ &= (a_n^2 + b_n)M^2, \end{aligned} \tag{13}$$

where

$$b_n = \sum_{k=1}^{n-1} \frac{1}{k^2},$$

we have

$$\mathrm{Var}(J_n) = b_nM^2.$$

Note that

$$\begin{aligned} E(L_n^2) &= E\left[\left(\sum_k l_k^{(n)}\right)^2\right] \\ &= nE(l_n^2) + n(n-1)E(l_nl_n'), \end{aligned}$$

where l_n and l_n' are two different randomly chosen external branches. It can be shown (APPENDIX A) that

$$nE(l_n^2) = 2\frac{(a_{n+1}-1)}{n-1}M^2,$$

$$n(n-1)E(l_nl_n') = \frac{2}{(n-1)(n-2)} \cdot \left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right)M^2.$$

Therefore,

$$\begin{aligned} \mathrm{Var}(L_n) &= E(L_n^2) - E^2(L_n) \\ &= 2\frac{(a_{n+1}-1)}{n-1}M^2 + \frac{2}{(n-1)(n-2)} \\ &\quad \cdot \left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right)M^2 - M^2 \\ &= c_nM^2, \end{aligned} \tag{14}$$

where $c_n = 1$ when $n = 2$ and when $n > 2$

$$c_n = 2\frac{na_n - 2(n-1)}{(n-1)(n-2)}.$$

Notice that

$$E(I_nL_n) = E(J_nL_n) - E(L_n^2),$$

$$E(I_n^2) = E(J_n^2) - 2E(J_nL_n) + E(L_n^2).$$

Thus, to calculate the variance of I_n and the covariance of I_n and L_n, one needs only to find $E(J_nL_n)$. Again, recurrent relationships can be used to show (APPENDIX A) that

$$E(J_nL_n) = \frac{n}{n-1}a_nM^2. \tag{15}$$

Therefore, we have

$$\begin{aligned} \mathrm{Cov}(I_n,L_n) &= E(I_nL_n) - E(I_n)E(L_n), \\ &= \left(\frac{1}{n-1}a_n - c_n\right)M^2, \end{aligned} \tag{16}$$

$$\begin{aligned} \mathrm{Var}(I_n) &= E(I_n^2) - E^2(I_n) \\ &= \left[a_n^2 + b_n - 2\left(\frac{n}{n-1}a_n - c_n\right) \right. \\ &\quad \left. + (c_n + 1) - (a_n - 1)^2\right]M^2, \\ &= \left(b_n - 2\frac{a_n}{n-1} + c_n\right)M^2. \end{aligned} \tag{17}$$

The numbers of mutations in external and internal branches: Let η_e and η_i be the total number of mutations in external and internal branches, respectively, and let $\eta = \eta_i + \eta_e$ be the total number of mutations that occurred in the entire genealogy of n

genes. Let the mutation rate per sequence per generation be μ and assume that the number of mutations that occur in a sequence in a period of l generations follows the Poisson distribution

$$\Pr(k \mid \mu l) = \frac{\exp(-\mu l)(\mu l)^k}{k!}.$$

Then the total number, η_e, of mutations in the external branches, given μL_n, follows the Poisson distribution

$$\Pr(\eta_e \mid \mu L_n) = \frac{\exp(-\mu L_n)(\mu L_n)^{\eta_e}}{\eta_e!}.$$

The expectation and variance of η_e are

$$\begin{aligned} E(\eta_e) &= E_{L_n}\{E(\eta_e \mid L_n)\} = E(\mu L_n) \\ &= \theta, \end{aligned} \tag{18}$$

$$\begin{aligned} \mathrm{Var}(\eta_e) &= E_{\{L_n\}}E(\eta_e^2 \mid L_n) - E^2(\eta_e) \\ &= \mu E(L_n) + \mu^2 E(L_n^2) - \mu^2 E(L_n) \\ &= \mu E(L_n) + \mu^2 \mathrm{Var}(L_n) \\ &= \theta + c_n\theta^2, \end{aligned} \tag{19}$$

where $\theta = 4N_e\mu$ and c_n is as defined above. It is interesting to note that

$$\lim_{n\to\infty} \mathrm{Var}(L_n) = M^2 \lim_{n\to\infty} c_n = 0.$$

This indicates that η_e is asymptotically Poisson distributed with parameter θ.

Similarly, we have

$$E(\eta) = E_{J_n}[E(\eta \mid J_n)] = \mu E(J_n) = a_n\theta, \tag{20}$$

$$\begin{aligned} \mathrm{Var}(\eta) &= E_{J_n}[E(\eta^2 \mid J_n)] - E^2(\eta) \\ &= \mu E(J_n) + \mu^2 \mathrm{Var}(J_n) \\ &= a_n\theta + b_n\theta^2. \end{aligned} \tag{21}$$

Equations 20 and 21 were also given by FU and LI (1993) and are the same as the expectation and variance of the number of segregating sites (K) in a sample of n sequences given by WATTERSON (1975) under the infinite site model. Similarly,

$$E(\eta_i) = E(\eta) - E(\eta_e) = (a_n - 1)\theta \tag{22}$$

$$\begin{aligned} \mathrm{Cov}(\eta_i,\eta_e) &= E_{I_n,L_n}[E(\eta_i\eta_e \mid L_n, I_n)] - E(\eta_i)E(\eta_e) \\ &= \mu^2 \mathrm{Cov}(I_n, L_n) \\ &= \left(\frac{a_n}{n-1} - c_n\right)\theta^2, \end{aligned} \tag{23}$$

and

$$\begin{aligned} \mathrm{Var}(\eta_i) &= \mathrm{Var}(\eta) - 2\,\mathrm{Cov}(\eta_i,\eta_e) - \mathrm{Var}(\eta_e) \\ &= a_n\theta + b_n\theta^2 - 2\left(\frac{a_n}{n-1} - c_n\right)\theta^2 - \theta - c_n\theta^2 \\ &= (a_n - 1)\theta + \left(b_n - 2\frac{a_n}{n-1} + c_n\right)\theta^2. \end{aligned} \tag{24}$$

Also,

$$\begin{aligned} \mathrm{Cov}(\eta,\eta_e) &= \mathrm{Var}(\eta_e) + \mathrm{Cov}(\eta_i,\eta_i) \\ &= \theta + c_n\theta^2 + \left(\frac{a_n}{n-1} - c_n\right)\theta^2 \\ &= \theta + \frac{a_n}{n-1}\theta^2. \end{aligned} \tag{25}$$

RELATIONSHIPS AMONG η_i, η_e AND Π_n

Let s_{ij} be the number of nucleotide differences between sequences i and j in the sample. Then, the mean number of pairwise differences for the n sequences is defined by

$$\Pi_n = \frac{2}{n(n-1)} \sum_{i<j} s_{ij}.$$

TAJIMA (1983, 1989) showed that the mean and variance of Π_n are

$$E(\Pi_n) = \theta, \tag{26}$$

$$\mathrm{Var}(\Pi_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \tag{27}$$

and the covariance of Π_n and η is

$$\mathrm{Cov}(\Pi_n,\eta) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2. \tag{28}$$

The covariance between Π_n and η_e or η_i can also be derived by considering recurrent relationships. It can be shown (APPENDIX B) that

$$\begin{aligned} \mathrm{Cov}(\Pi_n,\eta_e) &= 2\frac{n+1}{(n-1)^2} \\ &\quad \cdot \left(a_{n+1} - \frac{2n}{n+1}\right)\theta + \frac{1}{n-1}\theta^2, \end{aligned} \tag{29}$$

$$\begin{aligned} \mathrm{Cov}(\Pi_n,\eta_i) &= \mathrm{Cov}(\Pi_n,\eta) - \mathrm{Cov}(\Pi_n,\eta_e) \\ &= \left[1 - 2\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)\right]\theta \\ &\quad + \left(\frac{1}{2} + \frac{1}{n} - \frac{1}{n-1}\right)\theta^2. \end{aligned} \tag{30}$$

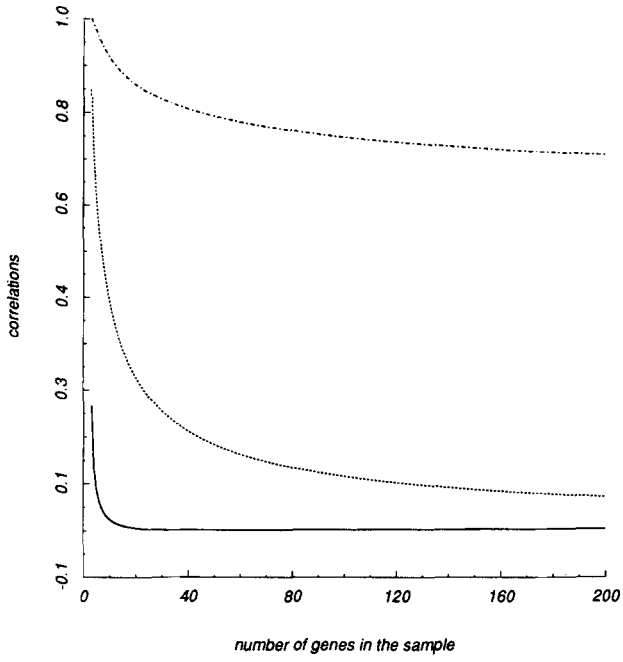


FIGURE 2.—Correlations. The solid, dotted and the dash-dotted curves are the correlations between η_e and η_i, between η_e and Π_n and between η and Π_n, respectively.

One can compute the correlations between various quantities considered here by using the formulas derived above and in the preceding sections. For example, the correlation between η_i and Π_n is

$$r_{\eta_i,\Pi_n} = \mathrm{Cov}(\eta_i,\Pi_n)/\sqrt{\mathrm{var}(\eta_i)\mathrm{Var}(\Pi_n)}.$$

which can be computed by using (24), (27) and (30).

Figure 2 shows the correlations between η_e and η_i, between η_e and Π_n and between η and Π_n. It is clear that η_e and η_i becomes almost independent when $n \geq 10$ while the correlation between η and Π_n remains strong for large n.

In the next two sections, the above results will be used to develop several test statistics. Tests with an outgroup will be considered first while tests with no outgroup will be considered later.

TEST STATISTICS WITH AN OUTGROUP

Let us first consider how to obtain the total number (η) of mutations and the number (η_e) of mutations in the external branches. Both η and η_e have to be inferred from sampled sequences. Inferring these two quantities by reconstructing the genealogy of the sampled genes is obviously the most accurate method. However, if infinite site model is assumed and an outgroup sequence is available, then there is a simpler way to obtain the two quantities. An outgroup is not part of the sample of n genes but a sequence whose common ancestor with the n genes in the sample is older than that of the n genes in the sample. Let s_i and e_i be, respectively, the total number of mutations and the number of mutations in the external branches in the sample at the ith site. Then s_i is equal to the number of different nucleotides minus one at site i among the n sequences and e_i is equal to the number of singleton nucleotides (excluding any singleton in the group sequence); a singleton nucleotide is a nucleotide that appears only once at the site among the sequences in the sample. Suppose that the sequences are m sites long. Then

$$\eta = \sum_i^m s_i \quad \text{and} \quad \eta_e = \sum_i^m e_i.$$

However, when there is no outgroup, it is difficult to infer accurately the number of external mutations. The test statistics of this kind will be considered later.

From the observed numbers of mutations in the internal branches and external branches, we have two unbiased estimates of θ, that is, $\eta_i/(a_n - 1)$ and η_e from Equations 18 and 22, respectively. We consider these two estimates because η_i and η_e become largely independent even when the sample size n is only moderately large. If neutrality of mutations does not hold, we expect the estimates $\eta_i/(a_n - 1)$ and η_e to be different. Therefore, the normalized difference between $\eta_i/(a_n - 1)$ and η_e can be used to test the null hypothesis. That is, we choose the test statistic as

$$\frac{\eta_i - (a_n - 1)\eta_e}{\sqrt{\mathrm{Var}(\eta_i - (a_n - 1)\eta_e)}} = \frac{\eta - a_n\eta_e}{\sqrt{\mathrm{Var}[\eta - a_n\eta_e]}},$$

where

$$\begin{aligned}\mathrm{Var}[\eta_i - (a_n - 1)\eta_e] &= \mathrm{Var}(\eta - a_n\eta_e)\\ &= \mathrm{Var}(\eta) + a_n^2\,\mathrm{Var}(\eta_e) - 2a_n\,\mathrm{Cov}(\eta,\eta_e)\\ &= \theta a_n(a_n - 1)\\ &\quad + \left[b_n + a_n(a_n - 2)c_n - 2a_n\left(\frac{a_n}{n-1} - c_n\right)\right]\theta^2 \qquad (31)\\ &= \theta a_n(a_n - 1)\\ &\quad + \left[b_n + a_n^2\left(c_n - \frac{2}{n-1}\right)\right]\theta^2.\end{aligned}$$

The normalization is intended to standardize the variance of the test statistic and hopefully bring the statistic close to the standard normal distribution.

In practice, the value of θ has to be estimated. As we have shown (FU and LI 1993), under the neutrality hypothesis, the estimate η/a_n of θ is asymptotically optimal. Therefore, we shall use it as the estimate of θ. As pointed out by TAJIMA (1989), $(\eta/a_n)^2$ is not an unbiased estimate of θ^2. The unbiased estimate of θ^2 is $\eta(\eta - 1)/(a_n^2 + b_n)$. With the replacement of θ and θ^2, respectively, by their unbiased estimates, the above

TABLE 1

Values of a_n, v_D, u_{D^*} and v_{D^*} as functions of sample size n

n	a_n	v_D	u_{D^*}	v_{D^*}	n	a_n	v_D	u_{D^*}	v_{D^*}	n	a_n	v_D	u_{D^*}	v_{D^*}
2	1.000	0.000	0.000	0.000	35	4.118	0.157	2.936	0.244	68	4.789	0.119	3.663	0.167
3	1.500	0.036	0.000	0.000	36	4.147	0.156	2.967	0.240	69	4.804	0.119	3.679	0.166
4	1.833	0.130	0.341	0.325	37	4.175	0.154	2.998	0.236	70	4.819	0.118	3.695	0.165
5	2.083	0.174	0.615	0.427	38	4.202	0.152	3.027	0.233	71	4.833	0.117	3.710	0.163
6	2.283	0.196	0.841	0.459	39	4.228	0.151	3.056	0.230	72	4.847	0.116	3.725	0.162
7	2.450	0.207	1.033	0.464	40	4.254	0.149	3.084	0.226	73	4.861	0.116	3.740	0.161
8	2.593	0.212	1.199	0.459	41	4.279	0.148	3.112	0.223	74	4.875	0.115	3.754	0.159
9	2.718	0.215	1.344	0.448	42	4.303	0.146	3.138	0.220	75	4.888	0.114	3.769	0.158
10	2.829	0.215	1.473	0.436	43	4.327	0.145	3.164	0.217	76	4.901	0.114	3.783	0.157
11	2.929	0.214	1.589	0.423	44	4.350	0.144	3.190	0.215	77	4.915	0.113	3.797	0.156
12	3.020	0.212	1.694	0.410	45	4.373	0.142	3.214	0.212	78	4.928	0.112	3.811	0.155
13	3.103	0.209	1.791	0.398	46	4.395	0.141	3.238	0.209	79	4.940	0.112	3.824	0.154
14	3.180	0.207	1.879	0.386	47	4.417	0.140	3.262	0.207	80	4.953	0.111	3.838	0.153
15	3.252	0.204	1.961	0.375	48	4.438	0.139	3.285	0.204	81	4.965	0.111	3.851	0.151
16	3.318	0.201	2.038	0.364	49	4.459	0.137	3.308	0.202	82	4.978	0.110	3.864	0.150
17	3.381	0.198	2.109	0.354	50	4.479	0.136	3.330	0.200	83	4.990	0.109	3.877	0.149
18	3.440	0.196	2.176	0.345	51	4.499	0.135	3.351	0.197	84	5.002	0.109	3.890	0.148
19	3.495	0.193	2.239	0.336	52	4.519	0.134	3.373	0.195	85	5.014	0.108	3.902	0.147
20	3.548	0.190	2.299	0.327	53	4.538	0.133	3.393	0.193	86	5.026	0.108	3.915	0.146
21	3.598	0.187	2.355	0.320	54	4.557	0.132	3.414	0.191	87	5.037	0.107	3.927	0.145
22	3.645	0.185	2.409	0.312	55	4.575	0.131	3.434	0.189	88	5.049	0.107	3.939	0.144
23	3.691	0.182	2.460	0.305	56	4.594	0.130	3.453	0.187	89	5.060	0.106	3.951	0.143
24	3.734	0.180	2.509	0.299	57	4.611	0.129	3.473	0.185	90	5.071	0.105	3.963	0.143
25	3.776	0.177	2.556	0.292	58	4.629	0.128	3.491	0.183	91	5.083	0.105	3.975	0.142
26	3.816	0.175	2.601	0.286	59	4.646	0.127	3.510	0.182	92	5.094	0.104	3.987	0.141
27	3.854	0.173	2.644	0.281	60	4.663	0.126	3.528	0.180	93	5.104	0.104	3.998	0.140
28	3.891	0.171	2.685	0.275	61	4.680	0.125	3.546	0.178	94	5.115	0.103	4.010	0.139
29	3.927	0.169	2.725	0.270	62	4.696	0.124	3.564	0.177	95	5.126	0.103	4.021	0.138
30	3.962	0.166	2.763	0.265	63	4.712	0.123	3.581	0.175	96	5.136	0.102	4.032	0.137
31	3.995	0.164	2.800	0.261	64	4.728	0.123	3.598	0.173	97	5.147	0.102	4.043	0.137
32	4.027	0.163	2.835	0.256	65	4.744	0.122	3.615	0.172	98	5.157	0.101	4.054	0.136
33	4.058	0.161	2.870	0.252	66	4.759	0.121	3.631	0.170	99	5.167	0.101	4.065	0.135
34	4.089	0.159	2.903	0.248	67	4.774	0.120	3.647	0.169	100	5.177	0.101	4.075	0.134

test statistic becomes

$$D = \frac{\eta - a_n\eta_e}{\sqrt{u_D\eta + v_D\eta^2}}, \tag{32}$$

where

$$v_D = 1 + \frac{a_n^2}{b_n + a_n^2}\left(c_n - \frac{n+1}{n-1}\right)$$

$$u_D = a_n - 1 - v_D .$$

The values of a_n and v_D are tabulated in Table 1 for $n = 2, \ldots, 100$. Note that a negative value of D means excess of mutations in external branches whereas a positive value means deficiency.

Unfortunately, like the test proposed by TAJIMA (1989), the test in (32) does not follow the standard normal distribution. The distribution is left skewed. Although the beta distribution may provide a better approximation than the standard normal distribution, it is not without problems. First, D is not a continuous variable, though approximately so when θ is very large. Second, the minimum and maximum values of D depend on θ. Third, there is also a problem in assuming a constant mean and variance, when inferring the parameters of the beta distribution. For example, TAJIMA (1989) assumed that the mean and variance of his test are 0 and 1, respectively, for all sample sizes but it is clear from his simulations that they depend on sample sizes. Because of these problems, we use the following computer simulation approach to determine the percentage points of the test. This approach should give more accurate results than using any approximation by standard distributions.

To determine the percentage points of the distribution of D for a given sample size n and a value of θ, we first simulate a large number (100,000) of samples each having exactly n sequences. The value of D is computed for each simulated sample and therefore the empirical distribution of D can be obtained. Properties related to the distribution such as percentage points can then be computed from the empirical distribution. For example, we compute the left tail percentage point $\alpha_{0.010}$, which is the maximum x value such that

$$\Pr(D \leq x) < 0.010,$$

TABLE 2

Percentage points of statistic D and D^* as functions of sample size n

	D						D^*					
n	$\alpha_{0.010}$	$\alpha_{0.025}$	$\alpha_{0.050}$	$\alpha_{0.950}$	$\alpha_{0.975}$	$\alpha_{0.990}$	$\alpha_{0.010}$	$\alpha_{0.025}$	$\alpha_{0.050}$	$\alpha_{0.950}$	$\alpha_{0.975}$	$\alpha_{0.990}$
4	−1.87	−1.68	−1.51	1.86	2.16	2.38	−0.87	−0.87	−0.87	1.89	2.08	2.19
5	−1.96	−1.77	−1.57	1.63	1.83	2.02	−1.26	−1.23	−1.20	1.57	1.68	1.77
6	−2.10	−1.88	−1.69	1.53	1.71	1.88	−1.54	−1.49	−1.43	1.46	1.55	1.62
7	−2.15	−1.90	−1.67	1.48	1.66	1.80	−1.75	−1.67	−1.57	1.37	1.46	1.56
8	−2.27	−1.97	−1.75	1.45	1.61	1.76	−1.93	−1.82	−1.67	1.34	1.43	1.51
9	−2.30	−2.04	−1.79	1.42	1.58	1.72	−2.07	−1.93	−1.74	1.32	1.40	1.49
10	−2.33	−2.11	−1.81	1.42	1.59	1.71	−2.19	−2.02	−1.79	1.30	1.38	1.48
11	−2.41	−2.18	−1.87	1.41	1.57	1.70	−2.30	−2.08	−1.86	1.27	1.37	1.47
12	−2.48	−2.12	−1.92	1.39	1.55	1.70	−2.39	−2.14	−1.87	1.26	1.36	1.47
13	−2.50	−2.19	−1.94	1.38	1.54	1.70	−2.49	−2.21	−1.91	1.29	1.37	1.47
14	−2.47	−2.26	−2.00	1.39	1.54	1.69	−2.54	−2.25	−1.92	1.28	1.36	1.47
15	−2.54	−2.18	−1.88	1.38	1.54	1.69	−2.61	−2.29	−1.93	1.27	1.36	1.47
16	−2.58	−2.19	−1.86	1.37	1.55	1.70	−2.68	−2.34	−1.96	1.27	1.35	1.48
17	−2.65	−2.22	−1.89	1.36	1.54	1.68	−2.75	−2.39	−1.98	1.26	1.35	1.47
18	−2.58	−2.21	−1.89	1.38	1.54	1.68	−2.79	−2.41	−1.97	1.25	1.36	1.49
19	−2.58	−2.23	−1.89	1.38	1.53	1.69	−2.84	−2.41	−1.97	1.25	1.35	1.49
20	−2.63	−2.26	−1.91	1.37	1.53	1.69	−2.87	−2.43	−2.02	1.29	1.37	1.50
21	−2.62	−2.25	−1.92	1.37	1.53	1.69	−2.93	−2.47	−1.99	1.29	1.37	1.50
22	−2.65	−2.24	−1.92	1.36	1.53	1.69	−2.99	−2.47	−1.96	1.29	1.37	1.50
23	−2.67	−2.25	−1.91	1.36	1.53	1.70	−3.02	−2.50	−1.95	1.29	1.37	1.50
24	−2.70	−2.27	−1.92	1.35	1.54	1.70	−3.04	−2.51	−1.96	1.28	1.37	1.50
25	−2.72	−2.29	−1.95	1.35	1.54	1.70	−3.08	−2.52	−1.95	1.28	1.38	1.51
26	−2.75	−2.31	−1.96	1.37	1.54	1.70	−3.09	−2.51	−1.94	1.28	1.38	1.52
27	−2.77	−2.33	−1.98	1.36	1.54	1.70	−3.11	−2.50	−1.92	1.27	1.38	1.52
28	−2.79	−2.34	−2.01	1.36	1.54	1.70	−3.17	−2.55	−1.95	1.27	1.38	1.52
29	−2.79	−2.35	−2.04	1.36	1.54	1.71	−3.17	−2.52	−1.96	1.27	1.38	1.54
30	−2.75	−2.37	−2.06	1.37	1.54	1.71	−3.18	−2.53	−1.91	1.27	1.39	1.54
32	−2.81	−2.38	−1.95	1.38	1.53	1.71	−3.25	−2.55	−1.94	1.32	1.40	1.54
34	−2.78	−2.32	−1.94	1.37	1.55	1.72	−3.23	−2.50	−1.96	1.31	1.40	1.55
36	−2.77	−2.33	−1.94	1.37	1.55	1.72	−3.28	−2.52	−2.00	1.31	1.41	1.55
38	−2.77	−2.32	−1.94	1.36	1.55	1.72	−3.29	−2.50	−2.05	1.31	1.40	1.57
40	−2.81	−2.35	−1.96	1.38	1.55	1.73	−3.33	−2.51	−1.86	1.31	1.42	1.58
42	−2.79	−2.34	−1.94	1.37	1.55	1.73	−3.34	−2.49	−1.88	1.30	1.42	1.57
44	−2.81	−2.36	−1.96	1.37	1.55	1.73	−3.35	−2.50	−1.86	1.30	1.42	1.59
46	−2.81	−2.36	−1.95	1.37	1.55	1.74	−3.40	−2.51	−1.84	1.30	1.44	1.59
48	−2.82	−2.38	−1.95	1.37	1.56	1.75	−3.37	−2.47	−1.87	1.30	1.44	1.60
50	−2.83	−2.39	−1.96	1.37	1.56	1.75	−3.38	−2.45	−1.88	1.30	1.44	1.61
55	−2.87	−2.45	−1.95	1.39	1.57	1.76	−3.34	−2.41	−1.87	1.31	1.46	1.62
60	−2.90	−2.39	−1.95	1.39	1.57	1.76	−3.41	−2.41	−1.90	1.34	1.46	1.63
65	−2.90	−2.39	−1.95	1.38	1.58	1.76	−3.39	−2.44	−1.87	1.34	1.47	1.64
70	−2.85	−2.35	−1.93	1.39	1.59	1.78	−3.27	−2.36	−1.19	1.33	1.48	1.66
75	−2.89	−2.38	−1.94	1.40	1.59	1.78	−3.32	−2.34	−1.89	1.33	1.49	1.67
80	−2.88	−2.35	−1.92	1.40	1.59	1.78	−3.22	−2.33	−1.91	1.33	1.50	1.68
85	−2.92	−2.39	−1.96	1.40	1.59	1.78	−3.40	−2.35	−1.88	1.33	1.50	1.68
90	−2.88	−2.35	−1.92	1.39	1.60	1.80	−3.27	−2.30	−1.91	1.33	1.51	1.70
95	−2.91	−2.37	−1.95	1.40	1.61	1.80	−3.19	−2.30	−1.94	1.37	1.52	1.70
100	−2.91	−2.36	−1.95	1.41	1.61	1.81	−3.27	−2.33	−1.90	1.37	1.53	1.71

and the right tail percentage point $\alpha_{0.950}$, which is the minimum x value such that

$$\Pr(D \geq x) < 0.050.$$

In principle, one should compare the observed D value with the distribution of D with the same θ value as that of the population from which the sequences are drawn. However, in practice, the θ value is unknown. The most appropriate distribution to compare with is perhaps the one with $\theta = \eta/a_n$, but tabulation of percentage points for many values of θ is not feasible. Instead, we choose to present conservative percentage points (Table 2). The actual probabilities corresponding to these percentage points can not be larger than the nominal level as long as the actual θ falls into the interval [2, 20]. We choose this interval because it should cover most of the situations of practical importance. To be precise, the percentage points for the

left tail is chosen by minimizing the percentage points over the above mentioned interval of θ and the right tail percentage points are chosen by maximizing the percentage points over the above mentioned interval of θ. Figure 3, a and b, shows examples of the variation among percentage points for different values of θ and n. It is clear that the percentage points corresponding to the left tail of the distribution of D are relatively constant over the range of values of θ, but those corresponding to the right tail are quite variable. This implies that the significant level based on a left tail percentage point will be very close to the nominal significant level but that based on a right tail percentage point might be too conservative. Interestingly, similar simulation on TAJIMA's test T showed that the right tail of statistic T is also very variable and the right tail percentage points in Table 2 of TAJIMA (1989) are too conservative. The conservative percentage points of statistic T generated by using computer simulation can be obtained from the authors upon request.

Note that both two-sided and one-sided tests can be conducted using the statistic D. For example, if the observed D is -2.20 with 10 genes ($n = 10$), then the result is significant at the 2.5% level if the test is one-sided and is significant at the 5% level if the test is two-sided.

TEST STATISTICS WITH NO OUTGROUP

When there is no outgroup available, the number of singletons (η_s) may overestimate the number of mutations in the external branches. To illustrate this point, consider the genealogy in Figure 4a. The mutations on branch a are all singletons under the infinite site model, though branch a is an internal branch under the definition used in this paper. For this reason, the percentage points for the distribution of D no longer gives accurate percentage points when η_e is replaced with η_s. A new test constructed using the moments of η_s and $\eta - \eta_s$ is required. We now consider such a test.

Since we are considering only bifurcating trees, there are exactly two branches leading to the root of a genealogy. Let κ be the number of external branches leading to the root of the tree. Then for $n > 2$, κ is either 1 or 0. For example, $\kappa = 1$ for the Figure 4a and 0 for Figure 4b. TAJIMA (1983) showed $\Pr(\kappa = 1) = 2/(n - 1)$. Consider the case of $\kappa = 1$. Let the number of mutations on the internal branch (branch a) be ξ_i and the number of mutations on the external branch (branch b) occurring during the coalescent time t_2 be ξ_e. Then, $E(\xi_i) = E(\xi_e) = 2N_e\mu = \theta/2$.

In general, without any outgroup we cannot locate the root of the tree (*i.e.*, we do not know whether $\kappa =$ 1 or 0) and so the quantity η_s should be defined as

$$\eta_s = \eta_e + \zeta,$$

where

$$\zeta = \begin{cases} 0 & \text{if } \kappa = 0, \\ \xi_i & \text{if } \kappa = 1. \end{cases} \tag{33}$$

Because $\Pr(\kappa = 1) = 2/(n - 1)$ (TAJIMA 1983), we have

$$E(\zeta) = \frac{2}{n-1} E(\xi_i) = \frac{\theta}{n-1},$$

and

$$E(\eta_s) = \theta + \frac{\theta}{n-1} = \frac{n}{n-1}\theta.$$

Note that the same result was obtained by TAJIMA (1989) using the infinite allele model for each site, though each site has at most four alleles. Moreover,

$$E(\eta_i - \zeta) = \left(a_n - \frac{n}{n-1}\right)\theta.$$

This suggests that the normalized differences between $\eta_i - \zeta$ and η_s can be used as the test statistic. However, because

$$\frac{\eta_i - \zeta}{a_n - n/(n-1)} - \frac{\eta_s}{n/(n-1)}$$

$$= \left[\left(a_n - \frac{n}{n-1}\right)\frac{n}{n-1}\right]^{-1}\left(\frac{n}{n-1}\eta - a_n\eta_s\right),$$

we use the following test statistic

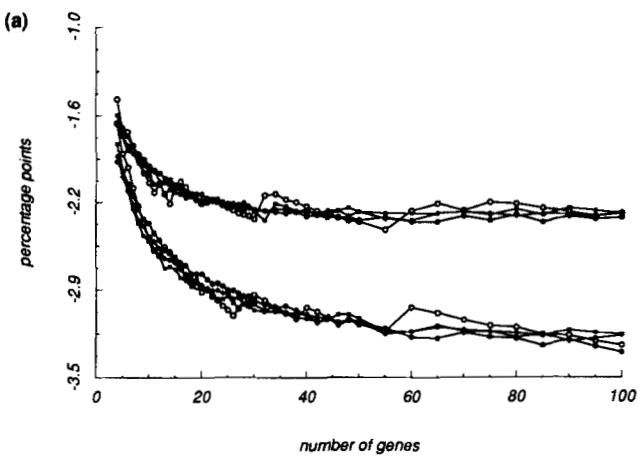
$$\frac{\left(\frac{n}{n-1}\right)\eta - a_n\eta_s}{\sqrt{\mathrm{Var}\left(\frac{n}{n-1}\eta - a_n\eta_s\right)}},$$

where

$$\mathrm{Var}\left(\frac{n}{n-1}\eta - a_n\eta_s\right) = u_n'\theta + v_n'\theta^2.$$

The derivation of the variance and the values of u_n' and v_n' are given in APPENDIX C. Following the same approach as in the derivation of D, and replacing θ by its estimate η/a_n and θ^2 by $\eta(\eta - 1)/(a_n^2 + b_n)$, we have the following test statistic:

$$D^* = \frac{\left(\frac{n}{n-1}\right)\eta - a_n\eta_s}{\sqrt{u_{D^*}\eta + v_{D^*}\eta^2}}$$



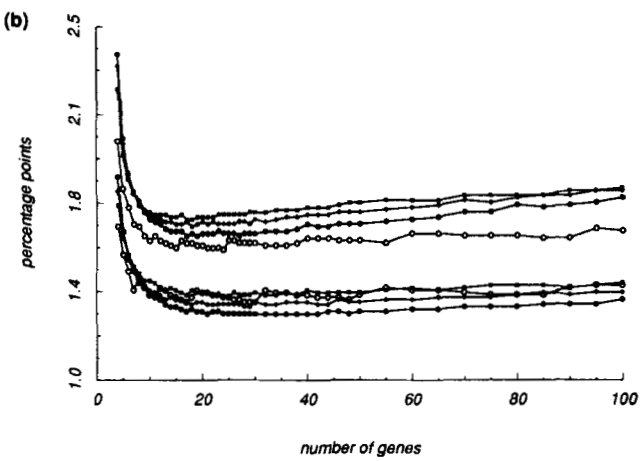


FIGURE 3.—Examples of percentage points as functions of θ and n. (a) 1% (bottom group) and 5% (top groups) left tail percentage points, and (b) 1% (bottom group) and 5% (top groups) right tail percentage points of the test statistic D. In each group of curves, lines with circles corresponds to $\theta = 2$, lines with solid circles to $\theta = 5$, lines with diamonds to $\theta = 10$ and lines with solid squares to $\theta = 20$. Each point in a curve is from an empirical distribution generated from 100,000 simulated samples.

where

$$v_{D^*} = \left[\left(\frac{n}{n-1}\right)^2 b_n + a_n^2 d_n - 2\frac{n a_n (a_n+1)}{(n-1)^2}\right] \Big/ (a_n^2 + b_n),$$

$$u_{D^*} = \frac{n}{n-1}\left(a_n - \frac{n}{n-1}\right) - v_{D^*},$$

in which d_n is defined by (46) in APPENDIX C. For convenience of application, the values of u_{D^*} and v_{D^*} are presented in Table 1. Again, we use computer simulations to determine the conservative percentage points of the distribution of this statistic over the interval [2, 20] of θ. These percentage points are given in Table 2. Detailed analysis showed that the 5% and 2.5% percentage points for the left tail are very stable over the range of θ but the 1% and 0.5% percentage points for the left tail are quite variable when the sample size (n) is larger than 20. The percentage points for the right tail of the distribution have similar variation as that of the statistic D. Therefore, the 1% and 0.5% left tail significance levels in Table 2 might be quite conservative when the sample size is larger than 20 while all the right tail significant levels are conservative.

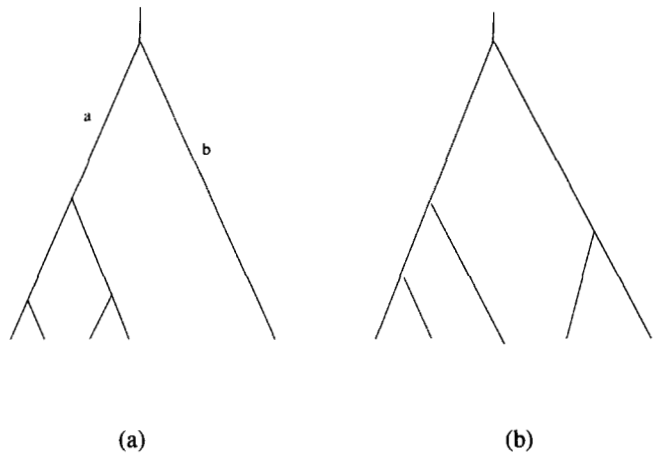


FIGURE 4.—Examples of genealogy. (a) One of the two branches leading to the root of the tree is an external branch ($\kappa = 1$) and (b) Both branches leading to the root of the tree are internal ($\kappa = 0$).

It should be pointed out that since our test intends to compare the mutations in the recent past with those of relatively remote past, it is always better to use an outgroup whenever available. The outgroup should be from closely related population or species to avoid the complication caused by parallel and back mutations.

OTHER TESTS

There are other tests that can be constructed using the results derived in this paper. For example, one can use the normalized difference between η_e and Π_n as a test statistic

$$\frac{\Pi_n - \eta_e}{\sqrt{\mathrm{Var}(\Pi_n - \eta_e)}},$$

where

$$\mathrm{Var}(\Pi_n - \eta_e) = \mathrm{Var}(\Pi_n) + \mathrm{Var}(\eta_e) - 2\mathrm{Cov}(\eta_e, \Pi_n)$$
$$= u\theta + v\theta^2,$$

and

$$u = 1 + \frac{n+1}{3(n-1)} - 4\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)$$

$$v = c_n + \frac{2(n^2+n+3)}{9n(n-1)} - \frac{2}{n-1}.$$

Replacing θ and θ^2, respectively, by η/a_n and $\eta(\eta-1)/(a_n^2 + b_n)$, we have the following test:

$$F = \frac{\Pi_n - \eta_e}{\sqrt{u_F\eta + v_F\eta^2}},$$

where

$$v_F = \left[c_n + \frac{2(n^2+n+3)}{9n(n-1)} - \frac{2}{n-1}\right]/(a_n^2 + b_n),$$

$$u_F = \left[1 + \frac{n+1}{3(n-1)} - 4\frac{n+1}{(n-1)^2} \cdot \left(a_{n+1} - \frac{2n}{n+1}\right)\right]/a_n - v_F.$$

A test analogous to F using η_s can also be constructed. The test statistic is

$$\frac{\frac{n-1}{n}\eta_s - \Pi_n}{\sqrt{\mathrm{Var}\left(\frac{n-1}{n}\eta_s - \Pi_n\right)}},$$

where

$$\mathrm{Var}\left(\Pi_n - \frac{n-1}{n}\eta_s\right) = \mathrm{Var}(\Pi_n) + \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(\eta_s) - 2\frac{n-1}{n}\mathrm{Cov}(\eta_s,\Pi_n).$$

From the results in APPENDIX C, this variance can be calculated. After replacing θ and θ^2, respectively, by η/a_n and $\eta(\eta-1)/(a_n^2+b_n)$, we arrive at the following test statistic:

$$F^* = \frac{\Pi_n - \frac{n-1}{n}\eta_s}{\sqrt{u_{F^*}\eta + v_{F^*}\eta^2}},$$

where

$$v_{F^*} = \left[d_n + \frac{2(n^2+n+3}{9n(n-1)} - 2\frac{1}{n-1}\left(4b_n - 6 + \frac{8}{n}\right)\right] \Big/ (a_n^2 + b_n),$$

$$u_{F^*} = \left[\frac{n}{n-1} + \frac{n+1}{3(n-1)} - 2\frac{2}{n(n-1)} + 2\frac{n+1}{(n-1)^2} \cdot \left(a_{n+1} - \frac{2n}{n+1}\right)\right]/a_n - v_{F^*}.$$

The conservative percentage points of both F and F^* are given in Table 4 and for the convenience of computation, the values of u_F, v_F, u_{F^*} and v_{F^*} are given in Table 3. As noted above, the idea of our tests is to compare mutations occurred in the remote past with mutations in the recent past, and so the statistic F is preferred over F^* if an outgroup is available.

AN EXAMPLE

We now use the *Adh* gene sequence data of 12 individuals from *Drosophila yakuba* in McDONALD and KREITMAN (1991) to illustrate the use of the tests developed in the paper. Because outgroup sequences from *Drosophila simulans* and *Drosophila melanogaster* are available, η_e can be determined. For this data set, we have $n = 24$ because each individual represents two sequences and

$$\Pi_n = 3.16, \quad \eta = 18, \quad \eta_e = 9, \quad \eta_s = 10.$$

From these quantities, the values of test statistics T, D, D^*, F, and F^* can be computed. For example, from Table 1, one can find that for $n = 24$, $a_n = 3.734$ and $v_D = 0.1797$. Therefore, $u_D = a_n - 1 - v_D = 2.555$. Thus

$$D = \frac{\eta - a_{24}\eta_e}{\sqrt{u_D\eta + v_D\eta^2}} = \frac{18 - 3.734 \times 9}{\sqrt{18(2.555 + 0.1797 \times 18)}} = \frac{-15.606}{\sqrt{10.421}} = -1.529.$$

Comparing this value with the left tail percentage points (Table 2), we can see that it is not significant at 5% even for a one sided test. The values of the other tests can be found to be

$$T = -1.243, \quad D^* = -1.558, \quad F = -1.735, \quad \text{and} \quad F^* = -1.710.$$

Comparing these values of tests with corresponding tables of percentage points, it is also found that none of the tests is significant, though they all show excess of external mutations.

As noted before, a more rigorous analysis can be done by comparing these values with percentage points from a narrower range of θ values in the neighborhood of $\eta/a_n = 4.82$. The variance of this estimate can be obtained from (21) by replacing θ and θ^2, respectively, by η/a_n and $\eta(\eta-1)/(a_n^2+b_n)$. The estimated standard error of the estimate η/a_n is 1.885. We therefore choose to conduct a detailed analysis assuming that the true θ is within the range $[\eta/a_n - 2 \times \text{s.e.}, \eta/a_n + 2 \times \text{s.e.}] = [1.05, 8.59]$. Table 5 summarizes the results of the analysis. Test F^* gives the smallest mean probability while TAJIMA's test T gives the largest mean probability. Although this does not necessarily indicate larger powers for our tests, our tests seem to be promising.

TABLE 3

Values of u_F, v_F, u_{F^*} and v_{F^*} as functions of sample size n

n	u_F	v_F	u_{F^*}	v_{F^*}	n	u_F	v_F	u_{F^*}	v_{F^*}	n	u_F	v_F	u_{F^*}	v_{F^*}
2	0.000	0.000	0.000	0.000	35	0.245	0.017	0.233	0.023	68	0.232	0.012	0.227	0.014
3	0.206	0.016	0.000	0.000	36	0.245	0.017	0.233	0.022	69	0.232	0.012	0.227	0.014
4	0.219	0.043	0.067	0.070	37	0.244	0.016	0.233	0.022	70	0.232	0.011	0.227	0.014
5	0.229	0.047	0.109	0.083	38	0.244	0.016	0.233	0.021	71	0.232	0.011	0.227	0.014
6	0.236	0.045	0.138	0.081	39	0.243	0.016	0.233	0.021	72	0.231	0.011	0.226	0.014
7	0.241	0.043	0.159	0.077	40	0.243	0.016	0.233	0.021	73	0.231	0.011	0.226	0.014
8	0.244	0.040	0.174	0.071	41	0.242	0.015	0.233	0.020	74	0.231	0.011	0.226	0.013
9	0.247	0.038	0.186	0.066	42	0.242	0.015	0.233	0.020	75	0.230	0.011	0.226	0.013
10	0.248	0.036	0.195	0.061	43	0.242	0.015	0.232	0.020	76	0.230	0.011	0.225	0.013
11	0.250	0.034	0.202	0.057	44	0.241	0.015	0.232	0.019	77	0.230	0.011	0.225	0.013
12	0.250	0.032	0.207	0.054	45	0.241	0.015	0.232	0.019	78	0.229	0.011	0.225	0.013
13	0.251	0.031	0.212	0.050	46	0.240	0.014	0.232	0.019	79	0.229	0.011	0.225	0.013
14	0.251	0.029	0.215	0.047	47	0.240	0.014	0.232	0.018	80	0.229	0.011	0.225	0.013
15	0.252	0.028	0.219	0.045	48	0.240	0.014	0.232	0.018	81	0.229	0.011	0.224	0.013
16	0.252	0.027	0.221	0.043	49	0.239	0.014	0.231	0.018	82	0.228	0.011	0.224	0.013
17	0.252	0.026	0.223	0.041	50	0.239	0.014	0.231	0.018	83	0.228	0.011	0.224	0.012
18	0.251	0.025	0.225	0.039	51	0.238	0.014	0.231	0.017	84	0.228	0.010	0.224	0.012
19	0.251	0.024	0.226	0.037	52	0.238	0.013	0.231	0.017	85	0.227	0.010	0.223	0.012
20	0.251	0.024	0.228	0.036	53	0.238	0.013	0.231	0.017	86	0.227	0.010	0.223	0.012
21	0.251	0.023	0.229	0.034	54	0.237	0.013	0.230	0.017	87	0.227	0.010	0.223	0.012
22	0.250	0.022	0.230	0.033	55	0.237	0.013	0.230	0.016	88	0.227	0.010	0.223	0.012
23	0.250	0.022	0.230	0.032	56	0.237	0.013	0.230	0.016	89	0.226	0.010	0.223	0.012
24	0.250	0.021	0.231	0.031	57	0.236	0.013	0.230	0.016	90	0.226	0.010	0.222	0.012
25	0.249	0.021	0.232	0.030	58	0.236	0.013	0.229	0.016	91	0.226	0.010	0.222	0.012
26	0.249	0.020	0.232	0.029	59	0.235	0.013	0.229	0.016	92	0.226	0.010	0.222	0.012
27	0.249	0.020	0.232	0.028	60	0.235	0.012	0.229	0.015	93	0.225	0.010	0.222	0.012
28	0.248	0.019	0.233	0.027	61	0.235	0.012	0.229	0.015	94	0.225	0.010	0.222	0.012
29	0.248	0.019	0.233	0.026	62	0.234	0.012	0.229	0.015	95	0.225	0.010	0.221	0.011
30	0.247	0.018	0.233	0.026	63	0.234	0.012	0.228	0.015	96	0.225	0.010	0.221	0.011
31	0.247	0.018	0.233	0.025	64	0.234	0.012	0.228	0.015	97	0.224	0.010	0.221	0.011
32	0.246	0.018	0.233	0.024	65	0.233	0.012	0.228	0.015	98	0.224	0.010	0.221	0.011
33	0.246	0.017	0.233	0.024	66	0.233	0.012	0.228	0.014	99	0.224	0.010	0.221	0.011
34	0.246	0.017	0.233	0.023	67	0.233	0.012	0.227	0.014	100	0.224	0.010	0.220	0.011

DISCUSSION

As noted above, there are two types of free topologies in terms of the two branches leading to the root. The first type is that both branches are internal (*i.e.*, $\kappa = 0$) and is represented by Figure 4b. The other type is that one of the two branches is external (*i.e.*, $\kappa = 1$) and is represented by Figure 4a. Obviously, a tree with $\kappa = 1$ is likely to have more external mutations than a tree with $\kappa = 0$. This leads to an important question: should the information about the value of κ be used in a statistical test of the neutrality of mutations? The conditionality principle of inference [for example, see COX and HINKLEY (1974)] states that inference, particularly hypothesis testing, should be made by conditioning on the observed values of *ancillary variables*, that is, variables that are independent of whether or not the hypothesis being test is true. TAJIMA (1983) has shown that under selective neutrality $\Pr(\kappa = 1) = 2/(n - 1)$, where n is the sample size. In the presence of selection, this probability is likely to be different. Therefore, κ may not be an ancillary variable and one should be cautious in using a conditional test based on the value of κ.

However, if the conditional test statistics based on the value of κ are to be developed, the most appropriate way to construct such a test is to derive all the necessary expectations, variances and covariances conditioning on the value of κ and then substitute them into the statistic D. However, one can avoid such a tedious process by continuing to use the test statistic D with percentage points generated from genealogies of the same κ (say 1), though this is a less desirable approach. Two such tables, one for $\kappa = 0$ and the other for $\kappa = 1$, can be obtained from the authors upon request. The two conditional tests have similar patterns as the unconditional test D. In particular, the left tail percentage points are insensitive to the value of θ. The relationship among the unconditional test statistic D and conditional ones are

$$\Pr(D > \alpha) = \Pr(D > \alpha \mid \kappa = 0)\Pr(\kappa = 0) + \Pr(D > \alpha \mid \kappa = 1)\Pr(\kappa = 1)$$

The present study assumes no recombination. It is rather easy to see that recombination reduces the variance of the difference between any two estimates of θ from η, η_e, η_i, η_s or Π_n. For example, let us assume

TABLE 4

Percentage points of statistic F and F^* as functions of sample size n

	F						F^*					
n	$\alpha_{0.010}$	$\alpha_{0.025}$	$\alpha_{0.050}$	$\alpha_{0.950}$	$\alpha_{0.975}$	$\alpha_{0.990}$	$\alpha_{0.010}$	$\alpha_{0.025}$	$\alpha_{0.050}$	$\alpha_{0.950}$	$\alpha_{0.975}$	$\alpha_{0.990}$
4	−1.96	−1.78	−1.60	2.20	2.53	2.78	−.95	−.94	−.94	2.07	2.26	2.38
5	−2.11	−1.90	−1.71	1.91	2.15	2.36	−1.37	−1.35	−1.32	1.73	1.85	1.95
6	−2.29	−2.07	−1.87	1.79	2.00	2.19	−1.69	−1.64	−1.58	1.61	1.69	1.79
7	−2.37	−2.17	−1.91	1.71	1.91	2.11	−1.93	−1.84	−1.74	1.53	1.61	1.74
8	−2.52	−2.22	−1.95	1.68	1.88	2.07	−2.12	−2.01	−1.85	1.48	1.60	1.73
9	−2.56	−2.26	−1.99	1.64	1.84	2.05	−2.27	−2.12	−1.92	1.48	1.60	1.72
10	−2.60	−2.33	−2.01	1.62	1.83	2.04	−2.40	−2.21	−1.97	1.47	1.59	1.71
11	−2.65	−2.40	−2.06	1.60	1.81	2.03	−2.51	−2.29	−2.07	1.44	1.58	1.72
12	−2.71	−2.44	−2.10	1.59	1.80	2.02	−2.61	−2.33	−2.03	1.45	1.58	1.72
13	−2.73	−2.40	−2.13	1.58	1.79	2.01	−2.67	−2.39	−2.07	1.44	1.59	1.74
14	−2.73	−2.41	−2.12	1.57	1.79	2.02	−2.73	−2.42	−2.08	1.43	1.58	1.74
15	−2.75	−2.40	−2.12	1.56	1.78	2.01	−2.80	−2.46	−2.08	1.43	1.59	1.74
16	−2.78	−2.41	−2.08	1.55	1.77	2.01	−2.86	−2.50	−2.11	1.43	1.59	1.75
17	−2.80	−2.43	−2.11	1.55	1.78	2.01	−2.92	−2.55	−2.12	1.44	1.60	1.76
18	−2.80	−2.42	−2.09	1.54	1.77	2.01	−2.95	−2.55	−2.15	1.43	1.60	1.77
19	−2.79	−2.44	−2.09	1.54	1.76	2.00	−2.99	−2.56	−2.14	1.43	1.60	1.78
20	−2.84	−2.45	−2.08	1.54	1.77	2.01	−3.02	−2.57	−2.09	1.44	1.61	1.78
21	−2.83	−2.45	−2.09	1.53	1.77	2.02	−3.08	−2.60	−2.11	1.43	1.60	1.78
22	−2.81	−2.44	−2.09	1.54	1.77	2.01	−3.12	−2.60	−2.08	1.44	1.61	1.79
23	−2.85	−2.43	−2.09	1.53	1.76	2.02	−3.14	−2.61	−2.06	1.43	1.61	1.79
24	−2.84	−2.44	−2.08	1.53	1.77	2.02	−3.16	−2.62	−2.06	1.43	1.61	1.80
25	−2.86	−2.45	−2.07	1.53	1.76	2.02	−3.18	−2.62	−2.06	1.44	1.62	1.82
26	−2.88	−2.43	−2.07	1.53	1.76	2.02	−3.17	−2.61	−2.06	1.44	1.62	1.82
27	−2.90	−2.45	−2.07	1.52	1.76	2.02	−3.18	−2.58	−2.05	1.44	1.62	1.82
28	−2.91	−2.46	−2.07	1.52	1.77	2.03	−3.24	−2.62	−2.05	1.44	1.62	1.82
29	−2.91	−2.46	−2.05	1.52	1.76	2.03	−3.22	−2.60	−2.04	1.44	1.63	1.83
30	−2.88	−2.45	−2.05	1.52	1.76	2.04	−3.21	−2.59	−2.02	1.44	1.63	1.84
32	−2.86	−2.44	−2.05	1.53	1.76	2.03	−3.26	−2.60	−2.00	1.45	1.64	1.84
34	−2.84	−2.42	−2.03	1.52	1.76	2.03	−3.24	−2.54	−2.00	1.44	1.63	1.85
36	−2.87	−2.43	−2.02	1.51	1.77	2.04	−3.27	−2.54	−1.97	1.45	1.64	1.86
38	−2.86	−2.43	−2.00	1.52	1.77	2.05	−3.27	−2.51	−1.98	1.45	1.64	1.87
40	−2.86	−2.42	−2.01	1.51	1.76	2.04	−3.30	−2.53	−1.97	1.44	1.64	1.87
42	−2.85	−2.41	−1.99	1.51	1.78	2.06	−3.29	−2.50	−1.96	1.44	1.65	1.90
44	−2.84	−2.40	−1.99	1.51	1.76	2.04	−3.27	−2.48	−1.95	1.45	1.65	1.89
46	−2.84	−2.39	−1.97	1.52	1.78	2.06	−3.30	−2.49	−1.93	1.45	1.66	1.91
48	−2.83	−2.38	−1.97	1.51	1.78	2.07	−3.28	−2.45	−1.94	1.46	1.66	1.92
50	−2.83	−2.38	−1.97	1.51	1.78	2.08	−3.27	−2.43	−1.93	1.45	1.66	1.93
55	−2.83	−2.36	−1.95	1.51	1.77	2.07	−3.20	−2.41	−1.94	1.45	1.67	1.94
60	−2.84	−2.35	−1.94	1.51	1.78	2.09	−3.24	−2.37	−1.92	1.46	1.68	1.97
65	−2.81	−2.33	−1.94	1.51	1.78	2.09	−3.20	−2.36	−1.92	1.46	1.68	1.96
70	−2.79	−2.32	−1.93	1.51	1.78	2.11	−3.07	−2.34	−1.91	1.47	1.69	1.98
75	−2.76	−2.31	−1.92	1.50	1.78	2.11	−3.09	−2.32	−1.91	1.46	1.69	1.99
80	−2.76	−2.32	−1.92	1.51	1.78	2.10	−3.01	−2.32	−1.91	1.46	1.70	2.00
85	−2.78	−2.31	−1.92	1.51	1.79	2.12	−3.12	−2.32	−1.91	1.46	1.71	2.01
90	−2.77	−2.29	−1.89	1.52	1.79	2.13	−3.00	−2.30	−1.88	1.47	1.72	2.03
95	−2.80	−2.31	−1.91	1.52	1.80	2.13	−2.92	−2.32	−1.89	1.47	1.73	2.03
100	−2.77	−2.28	−1.89	1.51	1.81	2.15	−2.97	−2.30	−1.88	1.47	1.73	2.04

that a sequence with θ is divided into two independent (free recombination) half each with $\theta/2$. Then from (31)

$$\mathrm{Var}(\eta - a_n\eta_e) = 2\left[a_n(a_n-1)\theta/2 + \left[b_n + a_n^2\left(c_n - \frac{2}{n-1}\right)\right](\theta/2)^2\right]$$

$$< a_n(a_n-1)\theta + \left[b_n + a_n^2\left(c_n - \frac{2}{n-1}\right)\right]\theta^2,$$

where the right hand side of the inequality is the variance of the difference assuming no recombination. Thus, recombination tends to make our tests conservative.

TABLE 5

Achieved significance levels of five tests over the interval [1.05, 8.59] of θ for the *D. yakuba* data

Prob.	Min. prob	Max. prob.	Mean. prob.	$\theta = 4.82$
$\Pr(T \leq -1.243)$	0.093	0.106	0.100	0.101
$\Pr(D \leq -1.529)$	0.088	0.098	0.093	0.093
$\Pr(D^* \leq -1.558)$	0.075	0.102	0.085	0.084
$\Pr(F \leq -1.734)$	0.078	0.094	0.082	0.081
$\Pr(F^* \leq -1.734)$	0.070	0.100	0.078	0.073

The present study also assumes a constant population size and no migration. The effect of these factors should be investigated in the future. Here we notice that migration may introduce rare alleles into a population and that a population expansion may produce excess of rare alleles. Thus an excess of mutations in the external branches of a genealogy or in general a significant test value does not necessarily imply the presence of natural selection.

The major difference between TAJIMA's T test and our tests is that T uses the difference between η and Π_n whereas our tests use wither the difference between η_i and η_e or the difference between η_e and Π_n. As can be seen from Figure 2, the correlation between η and Π_n is much stronger than that between η_i and η_e or η_e and Π_n. For this reason, our tests are likely to be more powerful than TAJIMA's test. However, a more careful comparison of the powers of TAJIMA's test and our tests needs to be made. The simplest way to compare the powers of various tests is to apply them to simulated samples that are generated without the assumption of neutrality of mutations. We intend to carry such studies in the future.

It should be emphasized that the present tests are for testing the hypothesis that all mutations in a DNA region are selectively neutral, but not for testing the neutral mutation hypothesis (KIMURA 1968). The latter assumes that the majority of mutations that can contribute significantly to the genetic variation at a locus are neutral or nearly neutral. This assumption is considerably weaker than the assumption that all mutations at the locus are neutral. Indeed, the neutral mutation hypothesis assumes that the most prevalent type of selection is purifying selection (KIMURA 1983). Thus, for example, even if all the rare variants at a locus are deleterious, the neutral mutation hypothesis still holds as long as the majority of the more common variants are selectively neutral; note that in this case, the hypothesis that all mutations are neutral does not hold. Some authors have failed to distinguish between the two hypotheses. For example, TAJIMA (1989) stated that his test was for testing the neutral mutation hypothesis, but like the present tests, it is for testing the assumption that all mutations are neutral.

This study was supported by National Institutes of Health grants.

LITERATURE CITED

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics*. Chapman & Hall, London.

FU, Y.-X., and W.-H. LI, 1993 Maximum likelihood estimation of population parameters. Genetics (in press).

HUDSON, R., 1982 Testing the constant-rate neutral allele model with protein sequence data. Evolution, **37:** 203–217.

HUDSON, R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, England.

KINGMAN, J., 1982 On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

MCDONALD, J., and M. KREITMAN, 1991 Adaptive protein evolution at *adh* locus in *Drosophila*. Nature **351:** 652–654.

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

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

WATTERSON, G., 1975 On the number of segregation sites. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: G. B. GOLDING

APPENDIX A. DERIVATIONS OF $E(l_n^2)$, $E(l_n l_n')$ AND $E(J_n L_n)$

To derive the expectation of a variable, it is often convenient to derive a function of the variable and obtain the required expectation by a simple transformation. In this and subsequent appendices, the function of a quantity whose expectation is to be derived will be denoted by g_n. The definition of g_n differs in different cases.

Let us consider $E(l_n^2)$ first and $g_n = n(n-1)l_n^2$. Then

$$\begin{aligned} E(g_n) &= 2(n-1)E(t_n^2) + (n-1)(n-2) \\ &\quad \cdot (l_{(n-1)}^2 + 2t_n l_{(n-1)} + t_n^2) \\ &= n(n-1)E(t_n^2) + E(g_{n-1}) + 2(n-1)(n-2) \\ &\quad \cdot E(t_n)E(l_{n-1}) \\ &= \frac{2M^2}{n(n-1)} + \frac{2(n-2)M^2}{n(n-1)} + E(g_{n-1}) \\ &= \frac{2M^2}{n} + E(g_{n-1}) = \cdots = 2(a_{n+1}-1)M^2. \end{aligned}$$

Therefore,

$$E(l_n^2) = \frac{E(g_n)}{n(n-1)} = \frac{2(a_{n+1}-1)}{n(n-1)} M^2. \qquad (34)$$

Next, we consider $E(l_n l_n')$. Let l_n and l_n' be two ran-

domly chosen different external branches. Then

$$l_n l'_n = \begin{cases} t_n^2 & P = \dfrac{2}{n(n-1)} \\ t_n(l_{(n-1)} + t_n) & P = \dfrac{2(n-2)}{n(n-1)} \\ (l'_{(n-1)} + t_n)t_n & P = \dfrac{2(n-2)}{n(n-1)} \\ (l_{(n-1)} + t_n)(l'_{(n-1)} + t_n) & P = \dfrac{(n-2)(n-3)}{n(n-1)}. \end{cases}$$

Therefore,

$$\begin{aligned} E(l_{(n)}l'_{(n)}) &= E(t_n^2) + \frac{2(n-2)}{n(n-1)}(2+n-3)E(l_{(n-1)})E(t_n) \\ &\quad + \frac{(n-2)(n-3)}{n(n-1)}E(l_{(n-1)}l'_{(n-1)}) \\ &= \cdot\frac{2}{n^2(n-1)^2}(1+(n-2)) \\ &\quad + \frac{(n-2)(n-3)}{n(n-1)}E(l_{(n-1)}l'_{(n-1)}) \\ &= \frac{2}{n^2(n-1)} + \frac{(n-2)(n-3)}{n(n-1)}E(l_{(n-1)}l'_{(n-1)}). \end{aligned}$$

Let

$$g_n = n(n-1)l_{(n)}l'_{(n)}.$$

Then

$$E(g_n) = \frac{2}{n} + \frac{n-3}{n-1}E(g_{n-1}). \tag{35}$$

We thus have $E(g_3) = 2/3$. In general

$$\begin{aligned} E(g_n) &= \frac{2}{n} + \frac{n-3}{n-1}E(g_{n-1}) \\ &= \frac{2}{n} + \frac{n-3}{n-1}\left(\frac{2}{n-1} + \frac{n-4}{n-2}E(g_{n-2})\right) \\ &= \frac{2}{n} + \frac{2(n-3)}{(n-1)^2} + \frac{(n-3)(n-4)}{(n-1)(n-2)}E(g_{n-2}) \\ &= \frac{2}{n} + \frac{2(n-3)}{(n-1)^2} + \frac{2}{(n-1)(n-2)} \\ &\quad \cdot\left(\frac{(n-3)(n-4)}{n-2} + \cdots + \frac{3\times 2}{4} + E(g_3)\right) \\ &= \frac{2}{(n-1)(n-2)}\sum_{k=1}^{n}\frac{(k-1)(k-2)}{k} \\ &= \frac{2}{(n-1)(n-2)}\left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right). \end{aligned}$$

Therefore,

$$E(l_n l'_n) = \frac{2}{(n-1)(n-2)}\left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right) \cdot \frac{1}{n(n-1)}. \tag{36}$$

Finally, we consider $E(J_nL_n)$. Since

$$J_{(n)} = J_{n-1} + nt_n$$

$$L_{(n)} = L_{n-1} + nt_n - l_{n-1},$$

we have

$$\begin{aligned} E(J_nL_n) &= \frac{n-2}{n-1}E(J_{n-1}L_{n-1}) + nE(J_{(n-1)})E(t_n) \\ &\quad + \frac{n(n-2)}{n-1}E(L_{(n-1)})E(t_n) + n^2E(t_n^2). \end{aligned}$$

Let $g_n = (n-1)J_nL_n$. After simplification, we have

$$\begin{aligned} E(g_n) &= E(g_{n-1}) + (a_n+1)M^2 = \cdots = \sum_{k=2}^{n}(a_k+1)M^2 \\ &= \sum_{k=1}^{n-1}\frac{n-k}{k}M^2 + (n-1)M^2 = na_nM^2. \end{aligned}$$

Therefore,

$$E(J_nL_n) = \frac{n}{n-1}a_nM^2. \tag{37}$$

APPENDIX B. COVARIANCE BETWEEN η_e AND Π_n

Let k_n be the number of nucleotide differences between two randomly selected sequences from a sample of n sequences and e_n be the number of mutations on a randomly selected external branch. Then

$$k_n = \begin{cases} \tau_i + \tau_j & \Pr = \dfrac{2}{n(n-1)}, \\ k_{n-1} + \tau_i + \tau_j & \Pr = 1 - \dfrac{2}{n(n-1)}, \end{cases}$$

where τ_i and τ_j are for two different branch segments but have the same time length t_n. Then the product of k_n and e_n has the recurrent relationship:

$$k_n e_n = \begin{cases} (\tau_i + \tau_j)\tau_i, & P_1 = \dfrac{2}{(n(n-1)}\dfrac{2}{n} \\ (\tau_i + \tau_j)(e_{n-1} + \tau_k), & P_2 = \dfrac{2}{n(n-1)}\dfrac{(n-2)}{n} \\ (k_{n-1} + \tau_i + \tau_k)\tau_i, & P_3 = \dfrac{4(n-2)}{n(n-1)}\dfrac{1}{n} \\ (k_{n-1} + \tau_i + \tau_k)\tau_j, & P_4 = \dfrac{4(n-2)}{n(n-1)}\dfrac{1}{n} \\ (k_{n-1} + \tau_i + \tau_k)(e_{n-1} + \tau_l), & P_5 = \dfrac{4(n-2)}{n(n-1)}\dfrac{(n-3)}{n} \\ (k_{n-1} + \tau_i + \tau_k)(e_{n-1} + \tau_k), & P_6 = \dfrac{4(n-2)}{n(n-1)}\dfrac{1}{n} \\ (k_{n-1} + \tau_k + t_l)\tau_i, & P_7 = \dfrac{(n-2)(n-3)}{n(n-1)}\dfrac{2}{n} \\ (k_n + \tau_k + \tau_l)(e_{n-1} + \tau_m), & P_8 = \dfrac{(n-2)(n-3)}{n(n-1)}\dfrac{(n-4)}{n} \\ (k_{n-1} + \tau_k + \tau_l)(e_{n-1} + \tau_k), & P_9 = \dfrac{(n-2)(n-3)}{n(n-1)}\dfrac{2}{n} \end{cases}$$

From the above recurrent relationship, we have

$$\begin{aligned} E(k_n e_n) = {} & E(k_{n-1}e_{n-1})(P_5 + P_6 \\ & + P_8 + P_9) \\ & + 2E(\tau)E(e_{n-1})(P_2 + P_5 + P_6 \\ & + P_8 + P_9) \\ & + E(\tau)E(k_{n-1})(P_3 + P_4 + P_5 + P_6 \\ & + P_7 + P_8 + P_9) \\ & + E(\tau^2)(P_1 + P_3 + P_6 \\ & + P_9) \\ & + E(\tau_i\tau_j)(P_1 \\ & + 2P_2 + P_3 + 2P_4 + 2P_5 + P_6 \\ & + 2P_7 + 2P_8 + P_9). \end{aligned}$$

Because

$$E(k_n) = \theta, \quad E(e_n) = \frac{\theta}{n}, \quad E(\tau) = \frac{\theta}{n(n-1)},$$

$$E(\tau^2) = \frac{\theta}{n(n-1)} + \frac{2\theta^2}{n^2(n-1)^2},$$

$$E(\tau_i\tau_j) = \frac{2\theta^2}{n^2(n-1)^2},$$

we can substitute these expectations and P's for those in Equation 38 and obtain

$$E(k_{(n)}\eta_{(n)}) = \frac{(n+1)(n-2)^2}{n^2(n-1)} E(k_{(n-1)}\eta_{(n-1)}) + \frac{2}{n^2(n-1)}\theta + \frac{n+2}{n^2(n-1)}\theta^2.$$

Note that

$$E(\Pi_n\eta_e) = nE(k_n e_n),$$

and let $g_n = nk_n e_n$. Then we have

$$E(g_n) = \frac{(n+1)(n-2)^2}{n(n-1)^2} E(g_{n-1}) + \frac{2}{n(n-1)}\theta + \frac{n+2}{n(n-1)}\theta^2.$$

After further simplification, we have

$$E(g_n) = a\theta + b\theta^2,$$

where

$$a = 2\frac{n+1}{(n-1)^2}\sum_{k=1}^{n}\frac{k-1}{k(k+1)} = 2\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)$$

$$b = \frac{n+1}{(n-1)^2}\sum_{k=1}^{n}\frac{(k-1)(k+2)}{k(k+1)} = \frac{n}{n-1}.$$

Therefore,

$$\begin{aligned} \mathrm{Cov}(\Pi_n, \eta_e) &= E(nk_n e_n) - E(k_n)E(ne_n) \\ &= 2\frac{n+1}{(n-1)^2}\left(a_{n+1} - \frac{2n}{n+1}\right)\theta + \frac{1}{n-1}\theta^2. \end{aligned} \tag{39}$$

APPENDIX C

Consider the genealogy of n genes conditioning on $\kappa = 1$, *i.e.* there is one external branch leading to the root of the three. Let l_n be the length of a randomly selected external branch that is not directly connected to the root and l'_n be the length of the external branch that is directly connected to the root. We then have the following recurrent relationship

$$l_n = \begin{cases} l_{n-1} + t_n & \Pr = \dfrac{n-3}{n-1} \\ t_n & \Pr = \dfrac{2}{n-1} \end{cases}$$

$$l'_n = l'_{n-1} + t_n$$

with initial condition $l_2 = l_1' = 0$. We thus have

$$E(l_n) = \frac{n-3}{n-1} E(l_{n-1}) + E(t_n).$$

Letting $h_n = (n-1)(n-2)l_n$, we have

$$E(h_n) = E(h_{n-1}) + \frac{n-2}{n} M = \sum_{k=2}^{n} \frac{k-2}{k} M$$

$$= [n - 1 - 2(a_{n+1} - 1)]M,$$

and

$$E(l_n') = \sum_{k=2}^{n} \frac{M}{k(k-1)} = \left(1 - \frac{1}{n}\right)M.$$

Let L_n be the total length of the external branches. Then

$$E(L_n \mid \kappa = 1) = \frac{n-1-2(a_{n+1}-1)}{n-2} M + \left(1 - \frac{1}{n}\right) M$$

$$= \left(2 - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right) M.$$

Therefore,

$$E(\eta_e \mid \kappa = 1) = \mu E(L_n \mid \kappa = 1)$$

$$= \left(2 - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right)\theta.$$

We now consider the variance of ζ, where ζ is defined as in Eq. (33)

$$\mathrm{Var}(\zeta) = E(\zeta^2) - E^2(\zeta)$$

$$= \frac{2}{n-1}(\mu E(t_2) + \mu^2 E(t_2^2)) - \frac{\theta^2}{(n-1)^2}$$

$$= \frac{1}{n-1}\theta + \frac{n-2}{(n-1)\theta^2}.$$

To derive the covariance between ζ and η_e, notice that

$$E(\eta_e\zeta) = E(\eta_e\zeta \mid \kappa = 1)\Pr(\kappa = 1)$$

$$= \mu^2 E([(n-1)l_n + l_n' - l_2' + l_2']t_2)\frac{n}{n-1}$$

$$= \frac{\theta^2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right) + \frac{\theta^2}{n-1},$$

Therefore,

$$\begin{aligned}\mathrm{Cov}(\eta_e,\zeta) &= E(\eta_e\zeta) - E(\eta_e)E(\zeta) \\ &= \frac{\theta^2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right) \\ &\quad + \frac{\theta^2}{n-1} - \theta\frac{\theta}{n-1} \\ &= \frac{\theta^2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right),\end{aligned} \tag{40}$$

$$\begin{aligned}\mathrm{Cov}(\eta,\zeta) &= \frac{2}{n-1}(E(\xi_i^2) + E(\xi,\xi_e) \\ &\quad + (a_n - 1)\theta^2/2) - \frac{a_n\theta^2}{n-1} \\ &= \frac{2}{n-1}(\theta/2 + \theta^2/2 \\ &\quad + \theta^2/2 + (a_n - 1)\theta^2/2) - \frac{a_n\theta^2}{n-1} \\ &= \frac{1}{n-1}(\theta + \theta^2).\end{aligned} \tag{41}$$

and

$$\begin{aligned}\mathrm{Cov}(\eta_i,\zeta) &= \mathrm{Cov}(\eta,\zeta) - \mathrm{Cov}(\eta_e,\zeta) \\ &= \frac{1}{n-1}(\theta + \theta^2) - \frac{1}{n-1} \\ &\quad \cdot \left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right)\theta^2.\end{aligned} \tag{42}$$

Note that

$$\begin{aligned}\mathrm{Var}(D^*) &= \left(\frac{n}{n-1}\right)^2 \mathrm{Var}(\eta) + a_n^2\mathrm{Var}(\eta_s) \\ &\quad - 2\frac{n}{n-1} a_n \mathrm{Cov}(\eta,\eta_s)\end{aligned} \tag{43}$$

where

$$\mathrm{Var}(\eta_s) = \mathrm{Var}(\eta_e) + \mathrm{Var}(\zeta) + 2\mathrm{Cov}(\eta_e,\zeta)$$

$$\mathrm{Cov}(\eta,\eta_s) = \mathrm{Cov}(\eta,\eta_e) + \mathrm{Cov}(\eta,\zeta)$$

Replacing each term on the right hand side of last two equations by its equation derived above, (for example, replacing cov(η,ζ) by Equation 40), we have

$$\mathrm{Var}(\eta_s) = \frac{n}{n-1}\theta + d_n\theta^2 \tag{44}$$

$$\mathrm{Cov}(\eta,\eta_s) = \frac{n}{n-1}\theta + \frac{a_n+1}{n-1}\theta^2 \tag{45}$$

where

$$d_n = c_n + \frac{n-2}{(n-1)^2} + \frac{2}{n-1} \cdot \left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right) \tag{46}$$

Putting these back to (43), we have

$$\mathrm{Var}(D^*) = u_n'\theta + v_n'\theta^2,$$

where

$$u_n' = \frac{na_n}{n-1}\left(a_n - \frac{n}{n-1}\right), \tag{47}$$

$$v_n' = \left(\frac{n}{n-1}\right)^2 b_n + a_n^2 d_n + -2\frac{na_n(a_n+1)}{(n-1)^2}. \tag{48}$$

We now consider the covariance between ζ and Π_n. Let k_n be the total time length between two randomly selected sequences (excluding the one that connects directly to the root). Then we have the recurrent relationship:

$$k_n = \begin{cases} 2t_n & \Pr = \dfrac{2}{(n-1)(n-2)} \\ k_{n-1} + 2t_n & \Pr = 1 - \dfrac{2}{(n-1)(n-2)}. \end{cases}$$

From this, we have

$$E(k_n) = \frac{n(n-3)}{(n-1)(n-2)} E(k_{n-1}) + 2E(t_n).$$

Let $h_n = (n-1)(n-2)k_n/2$. We then have

$$\begin{aligned} E(h_n) &= \frac{n}{n-2} E(h_{n-1}) + \frac{n-2}{n} M \\ &= \frac{n(n-1)}{(n-2)(n-3)} E(h_{n-2}) \\ &\quad + \frac{n(n-3)}{(n-1)(n-2)} M + \frac{n-2}{n} M = \cdots \\ &= \left[n(n-1)\sum_{i=2}^{n} \frac{i-2}{i^2(i-1)} + \frac{n-2}{n}\right] M \\ &= \left[n(n-1)\left(2b_n - 3 + \frac{1}{n-1}\right) + \frac{n-2}{n}\right] M. \end{aligned}$$

Therefore,

$$\begin{aligned} E(\Pi_n \mid \kappa = 1) &= \frac{2}{n(n-1)}\left[E(h_n)\mu + 2(n-1)\left(1 - \frac{1}{n}\right)\theta\right] \\ &= \left[4b_n - 6 + \frac{2}{n-1} \right. \\ &\quad \left. + \frac{2(n-2)}{n^2(n-1)} + 4\frac{n-1}{n^2}\right]\theta. \end{aligned}$$

The coefficient is always less than 1 when $n > 3$. This implies that when $\kappa = 1$, Π_n is an underestimate of θ. We finally have

$$\begin{aligned} \mathrm{Cov}(\zeta,\Pi_n) &= E(\Pi_n\zeta) - E(\Pi_n)E(\zeta) = \frac{1}{n-1} \\ &\quad \cdot \left[4b_n - 6 + \frac{2}{n-1} + \frac{2(n-2)}{n^2(n-1)} + 4\frac{n-2}{2n^2}\right]\theta^2 \\ &\quad + \frac{2}{n(n-1)}\frac{2}{n-1}[(n-1)E(\xi_i^2) \\ &\quad + (n-1)E(\xi_i\xi_e)] - \frac{\theta^2}{n-1} \\ &= \frac{1}{n-1}\left(4b_n - 6 + \frac{4}{n}\right)\theta^2 \\ &\quad + \frac{2}{n(n-1)}(\theta + 2\theta^2) - \frac{\theta^2}{n-1} \\ &= \frac{2}{n(n-1)}\theta + \frac{1}{n-1}\left(4b_n - 7 + \frac{8}{n}\right)\theta^2 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(\eta_s,\Pi_n) &= \mathrm{Cov}(\eta_e,\Pi_n) + \mathrm{Cov}(\zeta,\Pi_n) \\ &= \left[\frac{2}{n(n-1)} + 2\frac{n+1}{(n-1)^2}\right. \\ &\quad \left. \cdot \left(a_{n+1} - \frac{2n}{n+1}\right)\right]\theta \\ &\quad + \frac{1}{n-1}\left(4b_n - 6 + \frac{8}{n}\right)\theta^2. \end{aligned} \tag{49}$$