Copyright © 1989 by the Genetics Society of America

Gene Genealogy in Three Related Populations: Consistency Probability Between Gene and Population Trees

Naoyuki Takahata

National Institute of Genetics, Mishima, Shizuoka-Ken 411, Japan, and Center for Demographic and Population Genetics, The University of Texas Health Science Center, Houston, Texas 77225

Manuscript received November 2, 1988
Accepted for publication April 22, 1989

ABSTRACT

A genealogical relationship among genes at a locus (gene tree) sampled from three related populations was examined with special reference to population relatedness (population tree). A phylogenetically informative event in a gene tree constructed from nucleotide differences consists of interspecific coalescences of genes in each of which two genes sampled from different populations are descended from a common ancestor. The consistency probability between gene and population trees in which they are topologically identical was formulated in terms of interspecific coalescences. It was found that the consistency probability thus derived substantially increases as the sample size of genes increases, unless the divergence time of populations is very long compared to population sizes. Hence, there are cases where large samples at a locus are very useful in inferring a population tree.

THE nucleotide differences among genes at a locus drawn from a species contain useful information about how these genes evolved from a common ancestor. A genealogical relationship (gene tree) constructed from such nucleotide differences is a visual way of representing the evolutionary history of genes, through which not only the mechanisms of evolution of genes but also the evolutionary history of the species can be inferred. Furthermore, if orthologous (homologous) genes are drawn from different species or populations, the nucleotide differences can be used to infer the phylogenetic relationships of the species or populations (species or population tree).

However, even in the absence of gene flow, a gene tree does not necessarily show the same topological pattern as does a population tree (TAJIMA 1983; TAKAHATA and NEI 1985; NEIGEL and AVISE 1986; NEI 1987). This discordance stems from the fact that orthologous genes in different populations generally diverged much earlier than population splitting. Taking into account this possibility, NEI (1987) derived a simple formula for evaluating the probability that the topology of a tree for three orthologous genes, sampled from three different populations, is the same as that of the population tree. More recently, PAMILO and NEI (1988) extended the study of this problem to situations with more than three populations involved and those with more than one gene sampled from each population. They concluded that the consistency probability between gene and population trees becomes considerably smaller if internodal branches of the population tree are short and that this probability cannot be substantially increased by increasing the number of genes sampled from a locus.

In this paper, I shall address the same problem as did PAMILO and NEI (1988), and show that their conclusion, which seems rather discouraging to experimentalists, is largely due to the limited study of small sample sizes and the criterion they used. It is important to clearly distinguish two qualitatively different nodes in a gene tree. Each node (coalescence in the mathematical study of genealogy) (KINGMAN 1982) corresponds to a bifurcation of a gene in the reproduction process. A coalescence may be due to genes belonging to the same population or to different populations. These will be called intraspecific and interspecific coalescence, respectively. The occurrence of interspecific coalescence is a key event in a gene tree that can occur only before two populations involved have diverged from a common ancestor, and therefore it directly reflects population relatedness. Focusing on this event, I develop a theory relevant to the present problem and supplement the result with a simulation. It is then shown that sampling many genes from each population can indeed increase the consistency probability substantially, allowing us to correctly infer a population tree.

MODEL AND THEORY

The species considered here is monoecious and diploid. Generations are discrete and nonoverlapping, and for convenience they are counted backward chronologically from the present time. The species consists of three populations X, Y, and Z which se-

The publication costs of this article were partly defrayed by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

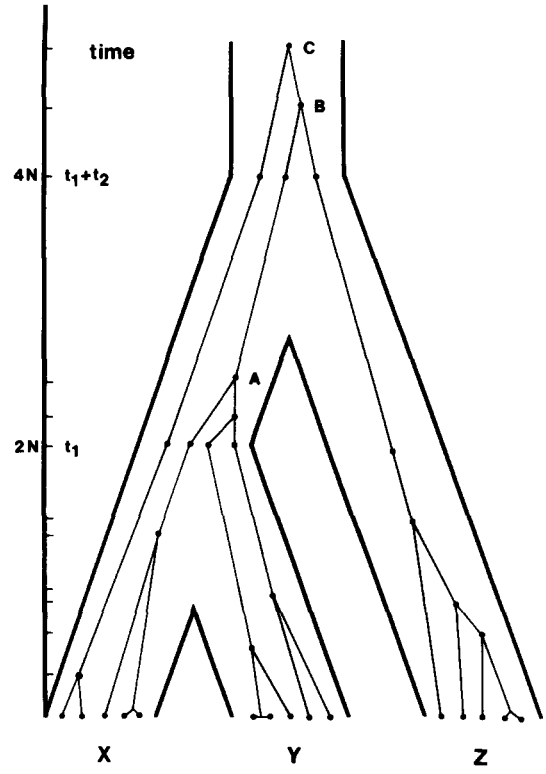

FIGURE 1.—Model of a population tree and a gene tree generated on a computer. *X*, *Y* and *Z* represent three different populations which diverged t_1 and $t_1 + t_2$ generations ago. Five genes were sampled from each population and $t_1 = t_2 = 2N$ were assumed. Dots and lines represent genes and ancestral lineages. Each node corresponds to a coalescence of genes. *A*, *B* and *C* stand for interspecific coalescences and all other nodes for intraspecific coalescences. In this simulation, there remained four ancestral genes from *X* and *Y* at t_1. Note that the probabilities that the first and the first two coalescences are intraspecific are 1/3 and 1/9, respectively.

quentially diverged from a common ancestral population t_1 and $t_1 + t_2$ generations ago (Figure 1). To be analytically accessible, two basic assumptions are made: neutrality (KIMURA 1968) and random mating in each population. An additional assumption is that a gene tree constructed from nucleotide differences is not subject to sampling errors which stem from comparisons of only a finite length of nucleotides. In other words, it is assumed that stochastic errors involved in a gene tree are solely due to random sampling drift.

Assume that each ancestral or descendant population consists of N selectively equivalent diploid individuals. Each descendant population at t_1 or $t_1 + t_2$ is formed by $2N$ gametes randomly sampled with replacement from the gamete pool of the parental population. To study the problem concerning the relationships between gene and population trees, we begin with the case of two populations X and Y which diverged t_1 generations ago. Suppose that we randomly draw r and s genes at a locus from X and Y, respectively, and trace back the ancestry until the time of the population splitting, t_1 generations ago. Orthologous genes in different populations evolve independently in the absence of gene flow and hence their common ancestor are found prior to the population splitting. Suppose that there existed $m(n)$ distinct ancestors of $r(s)$ sampled genes t_1 generations ago. Of interest here are the probability and time in which a particular type of coalescence occurs in the ancestral population. Previously TAKAHATA and NEI (1985) showed that the two closest genes in a sample can be used for estimating the divergence time of populations. However, they did not distinguish intraspecific and interspecific coalescences. Since intraspecific coalescence can occur in a descendant population, the two closest genes in a sample does not give precise information about the divergence time of populations. On the other hand, interspecific coalescence can occur only before the population splitting and set an upper bound of t_1. Thus it is interesting to know the probability and time in which interspecific coalescence occurs.

If there are $j = m + n$ distinct genes for a sample of size $r + s$ immediately after the population splitting, they must have been derived from a single common ancestor through $j - 1$ sequential bifurcations of genes in the ancestral population. In general, it is possible that more than two genes can be derived from a single parental gene, but this probability is very small for large N (KINGMAN 1982; TAJIMA 1983). Also, it is possible in our model of population splitting that j genes are derived from the same genes in the previous generation, t_1 ago. However, as discussed in TAKAHATA and NEI (1985), this probability is also very small and can be neglected. Hence we assume that only bifurcation of genes is allowed and that the number of distinct genes at t_1 is the sum of the number of distinct ancestors for a sample from two descendant populations.

Our first concern is with the probability, Q_{jk}, that the first $j - k$ coalescences occurring in the ancestral population of X and Y are intraspecific. In the ancestral population, there are two types of genes which are distinguished by whether their descendants belong to X or Y. In the present case, there are j genes at t_1 containing m genes of one type and n genes of a different type. When these j genes were derived from $j - 1$ genes by the first bifurcation, we randomly choose two from j genes and link them if they were of common type, or intraspecific. We continue this process $j - k$ times. Then Q_{jk} is the probability that we can trace successfully or intraspecifically back to k genes, or it is the probability that the number of coalescences back to the first interspecific coalescence is greater than k. To compute Q_{jk}, we define the probability that two genes randomly chosen from $m_0 (1 \leqq m_0 \leqq m)$ and $n_0 (1 \leqq n_0 \leqq n)$ genes are of identical type, and denote it by $P(m_0, n_0)$. $P(m, n) = 1$

corresponding to $Q_{jj} = 1$ ($j = m + n$), which reflects the boundary condition that there is no interspecific coalescence if there is no coalescence at all. Let k be $m_0 + n_0$.

Now we derive a recurrence equation for $P(m_0, n_0)$. We note that when a coalescence takes place among k genes, there are $k(k-1)/2$ different ways of pairing k genes. If m_0 genes are a result of all intraspecific coalescences, then there are $m_0(m_0-1)/2$ different ways of pairing for the next intraspecific coalescence. Likewise, there are $n_0(n_0-1)/2$ different ways of intraspecific coalescence for n_0 genes. Thus we have

$$P(m_0, n_0) = \frac{m_0(m_0+1)}{k(k+1)} P(m_0+1, n_0) + \frac{n_0(n_0+1)}{k(k+1)} P(m_0, n_0+1) \quad (1)$$

for $1 \leqq m_0 \leqq m$ and $1 \leqq n_0 \leqq n$, and otherwise $P(m_0, n_0) = 0$. The analytical solution of (1) may be found with boundary values

$$P(m_0, n) = \prod_{r=m_0}^{m-1} \frac{r(r+1)}{(r+n)(r+n+1)}, \quad (2)$$

and $P(m, n_0)$ which can be obtained from the right hand side of (2) if we exchange n and m and replace m_0 by n_0, respectively. For instance, in the case of $m = n = 2$, we have

$$P(1, 2) = P(2, 1) = 1/6, \quad P(1, 1) = 1/9. \quad (3)$$

Unfortunately, the general solution becomes rather untidy so that we use (1) numerically.

From (1), we can compute the probability, Q_{jk}, that the first $j-k$ coalescences are intraspecific by

$$Q_{jk} = \sum_r P(r, k-r) \quad (4)$$

where the summation is taken over r ranging from 1 or $k-n$ (whichever is larger) to $k-1$ or m (whichever is smaller). In relation to the example given in (3), we have

$$Q_{43} = P(1, 2) + P(2, 1) = 1/3, \quad Q_{42} = P(1, 1) = 1/9 \quad (5)$$

(Figure 1).

Some numerical results of (4) show that Q_{jk} decreases rather quickly as k decreases (Table 1), implying a rather high probability of interspecific coalescence occurrence. To see this in a slightly different way, we define D_{jk} as

$$D_{jk} = Q_{j,k+1} - Q_{jk} \quad (6)$$

for $1 \leqq k \leqq j-1$. This is the probability that the first interspecific coalescence occurs exactly when the number of distinct ancestors becomes k. For instance, $D_{j,j-1} = 2mn/[j(j-1)]$ ($j = m + n$) gives the probability that one of the m genes and one of the n genes are descended from a common ancestor. If $m = n = 1$, $D_{21} = 1$ since the coalescence is necessarily interspecific for two genes from different populations.

We are now at a position to evaluate the distribution of the time at which the first interspecific coalescence among j genes occurs in the ancestral population. We denote this random time by S_j, and define A as the random number of distinct ancestors of j genes at S_j. Thus Prob $(A = k) = D_{jk} = Q_{j,k+1} - Q_{jk}$ as mentioned above. Denote by τ_k the waiting time until k genes coalesce to $k-1$ genes, whose distribution is exponential with mean $4N/[k(k-1)]$ (KINGMAN 1982). For given A, we thus have

$$S_j = \tau_j + \tau_{j-1} + \ldots + \tau_{A+1} \quad (7)$$

and the continuous time version of the probability density is given by (4) in TAKAHATA and NEI (1985). In particular, the mean of S_j conditioned on $A = k$ is

$$E(S_j | A = k) = 4N\left(\frac{1}{k} - \frac{1}{j}\right) \quad (8)$$

(*e.g.*, KINGMAN 1982; TAJIMA 1983; TAVARÉ 1984), so that the unconditional mean of S_j becomes

$$E(S_j) = \sum_{k=1}^{j-1} E(S_j | A = k) D_{jk} = 4N\left(\sum_{k=1}^{j-1} \frac{1}{k} D_{jk} - \frac{1}{j}\right). \quad (9)$$

Thus $E(S_j) \geqq 4N/[j(j-1)]$ always holds true, implying that the unconditional mean of S_j (Table 1) is longer than that between the two closest genes in TAKAHATA and NEI (1985). By the same token, the unconditional probability density of S_j can be computed by

$$p(S_j) = \sum_{k=1}^{j-1} p(S_j | A = k) D_{jk} \quad (10)$$

where $p(S_j | A = k)$ is the probability density of S_j conditioned on $A = k$.

Now recall that the three populations X, Y and Z have a phylogenetical relationship as in Figure 1, and assume that both the topology and branch lengths are known. We are interested in the probability that a gene tree has the same topology as that of the population tree. It is to be noted, however, that when more than one gene is drawn from each population, the meaning of gene tree becomes equivocal because these genes often show different evolutionary relationships among different populations. PAMILO and NEI (1988) considered a composite gene tree for such cases, which is constructed by computing the average divergence time of genes taken over all pairwise comparisons between any pair of populations. It is not easy, however, to compute these average divergence times because they depend on the topology of the gene tree.

TABLE 1

Probability of no interspecific coalescence Q_{jk} and the expected waiting time until the first interspecific coalescence $E(S_j)$

			Q_{jk} $(2 \leq k \leq j)$, with $k =$						
$j = (m, n)$[a]	$E(S_j)$[b]	Ratio[c]	2	3	4	5	6	7	8
4 = (2, 2)	0.388	2.33	0.111	0.333	1				
4 = (3, 1)	0.500	3.00	0.167	0.500	1				
5 = (3, 2)	0.266	2.67	0.050	0.150	0.400	1			
5 = (4, 1)	0.400	4.00	0.100	0.300	0.600	1			
6 = (3, 3)	0.174	2.60	0.020	0.060	0.160	0.400	1		
6 = (4, 2)	0.200	3.00	0.027	0.080	0.200	0.467	1		
6 = (5, 1)	0.334	5.00	0.067	0.200	0.400	0.667	1		
7 = (4, 3)	0.126	2.64	0.010	0.029	0.074	0.181	0.429	1	
7 = (5, 2)	0.158	3.33	0.016	0.048	0.114	0.254	0.524	1	
7 = (6, 1)	0.286	6.00	0.048	0.143	0.286	0.476	0.714	1	
8 = (4, 4)	0.090	2.51	0.004	0.012	0.032	0.078	0.184	0.429	1
8 = (5, 3)	0.098	2.73	0.005	0.015	0.039	0.092	0.209	0.464	1
8 = (6, 2)	0.130	3.67	0.010	0.031	0.071	0.153	0.306	0.571	1
8 = (7, 1)	0.250	7.00	0.036	0.107	0.214	0.357	0.536	0.750	1

[a] Asymmetry for m and n, given $j = m + n$, increases the probability of intraspecific coalescence.
[b] $E(S_j)$ is measured in units of $2N$ generations.
[c] Ratio of $E(S_j)$ to expected waiting time until the first coalescence given by $2/[(j(j-1)]$.

This appears to be the main reason why PAMILO and NEI (1988) considered only two genes from each population. As sample size increases, a similar computation of the average divergence time seems extremely tedious, though not intractable. Another problem is concerned with the metric they used. As shown below, an average may not always be an appropriate measure because it overshadows minor relationships of genes that may be phylogenetically informative. It is thus necessary to reexamine the consistency probability between gene and population trees in more detail.

A key quantity is the probability that at least one interspecific coalescence occurs during the process in which j genes are derived from k distinct ancestors, given by

$$H_{jk} = 1 - Q_{jk}. \tag{11}$$

Suppose that we draw r and s genes from populations X and Y as before. Interspecific coalescence of these genes can occur prior to population splitting t_1 generations ago. But if it occurs only prior to the divergence between the common ancestral population of X and Y and population Z, $t_1 + t_2$ generations ago, the consistency between gene and population trees is nothing more than expected by mere chance (NEI 1987). Although such a coincidence must be taken into account in inferring the topology of a population tree, we will neglect it for the moment.

On the other hand, if at least one interspecific coalescence occurs between t_1 and $t_1 + t_2$, it becomes certain that populations X and Y are closer phylogenetically than Z, since genes from Z coalesce to those from X or Y necessarily before $t_1 + t_2$. In this situation, we say that a gene tree is consistent with a population tree. This probability is given by (11), provided that there existed $j = m + n$ distinct ancestors of r and s genes at t_1 and that there were $j - k$ coalescences between t_1 and $t_1 + t_2$. The distribution of the number of distinct ancestors k at t_2 in a stationary population for a sample of size j, $g_{jk}(t_2)$, was derived independently by TAVARÉ (1984), DONNELLY (1984), and TAKAHATA and NEI (1985) in which the relationship between $g_{jk}(t_2)$ and $p(S_j | A = k)$ in (10) was also given. Using (11) and $g_{jk}(t_2)$, we obtain the consistency probability between gene and population trees or the probability of at least one interspecific coalescence,

$$P = \sum_{k=1}^{j-1} g_{jk}(t_2) H_{jk}. \tag{12}$$

For $m = n = 1$ and thus $j = 2$, (12) becomes $g_{21}(t_2) = 1 - \exp(-t_2/(2N))$ since $H_{21} = 1$. For $m = n = 2$, it becomes

$$\begin{aligned} P &= g_{41}(t_2) + 8/9\, g_{42}(t_2) + 2/3\, g_{43}(t_2) \\ &= 1 - 1/5\, e^{-t_2/(2N)} \\ &\quad - 1/3\, e^{-3t_2/(2N)} - 7/15\, e^{-4t_2/N}. \end{aligned} \tag{13}$$

When $t_2/N \ll 1$, the ratio of P for $m = n = 2$ to that for $m = n = 1$ is about 5, implying a relatively high probability of occurrence of interspecific coalescence for $m > 1$ and $n > 1$.

We have assumed that the numbers of distinct genes m and n at t_1 are known. However, they are actually random numbers which again follow the same probabilistic law as in (12). Using $g_{rm}(t_1)$ and $g_{sn}(t_1)$, and recalling the independence of the genealogical processes in isolated populations X and Y, we finally obtain the consistency probability in terms of H_{jk} as

$$P = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{j-1} g_{rm}(t_1) g_{sn}(t_1) g_{jk}(t_2) H_{jk} \tag{14}$$

TABLE 2

Consistency probability, P, between gene and population trees computed from (14)

$(s, r)^a$	$t_1/(2N)^b$	$t_2/(2N)^b$ 0.05	0.5	5
(1, 1)	Any	0.049	0.394	0.993
(2, 2)	0.05	0.169	0.762	0.999
	0.5	0.118	0.628	0.999
	5	0.049	0.396	0.993
(5, 5)	0.05	0.604	0.989	1.0
	0.5	0.261	0.846	0.999
	5	0.050	0.399	0.993
(10, 10)	0.05	0.929	1.0	1.0
	0.5	0.371	0.921	1.0
	5	0.050	0.400	0.993

[a] Sample genes from population X and Y.

[b] Populations X, Y and their ancestor are assumed to have had a constant $2N$ genes in each population through time. The divergence time between X and Y is t_1 generations ago, and their common ancestor is assumed to have branched off from population Z, $t_1 + t_2$ generations ago.

where $j = m + n$. Numerical values of (14) for various values of parameters are given in Table 2.

SIMULATION AND RESULT

The genealogical process considered in the previous section was realized on a computer because an extension of PAMILO and NEI (1988) to the case of more than two genes from each population is very tedious and therefore the difference in their and our consistency probabilities is hard to evaluate analytically. A brief account of the simulation used is as follows.

Let r, s, and t be the numbers of genes sampled from the current populations X, Y, and Z, respectively. Let τ_{rst} be the holding time or waiting time in which a pair of genes in X, Y or Z coalesce to the most recent common ancestor. This time is exponentially distributed with mean $2/q_{rst}$ in units of $2N$ generations where

$$q_{rst} = r(r - 1) + s(s - 1) + t(t - 1). \quad (15)$$

Equation 15 is a consequence of the independent evolution of genes in different isolated populations and the assumption of sufficiently large N compared with sample size. This also implies that the probability that a coalescence occurs in X, Y and Z is respectively given by

$$P_X = r(r - 1)/q_{rst}, \quad P_Y = s(s - 1)/q_{rst}, \quad P_Z = t(t - 1)/q_{rst}. \quad (16)$$

To simulate this stochastic (death) process, we generate uniform and exponential random numbers. A uniform random number determines the population in which a coalescence occurs according to (16) and two additional numbers are used to determine a pair of coalescing genes in that population. An exponential random number determines how long the coalescence takes. This process reduces the number of distinct genes by one, and it is repeated until the number of distinct genes becomes one for the first time. However, when the cumulative coalescence time T taken over the repetition first exceeds T_1 or $T_1 + T_2$ [$T_1 = t_1/(2N)$ and $T_2 = t_2/(2N)$], it is necessary to take account of changes in population structure (Figure 1). When there remain m and n distinct genes at T_1 in the ancestral population of X and Y, and there remain k distinct genes in Z, (15) and (16) should be modified to

$$q_{jk} = j(j - 1) + k(k - 1), \quad j = m + n$$
$$P_{XY} = j(j - 1)/q_{jk}, \quad P_Z = k(k - 1)/q_{jk} \quad (17)$$

from T_1 to $T_1 + T_2$. Likewise a similar modification should be taken when T reaches $T_1 + T_2$ prior to which there exists only one panmictic population. In the simulation, it is also necessary to record each coalescence time and population in which the ancestral lineages of sampled genes reside. A simulation program which allows construction of the gene genealogy for an arbitrary sample size is available upon request.

In the case of $r = s = t = 1$ and $T_1 = T_2 = 1$, (14) or the argument about (13) leads to $P = 1 - \exp(-T_2) = 0.632$ while a simulation with 10^4 repeats yielded $P = 0.636$. In the case of $r = s = t = 2$ and $T_1 = T_2 = 1$, (14) predicts $P = 0.744$ while a simulation yielded $P = 0.742$. There is very close agreement between the theoretical and simulation results. On the other hand, if we compute the P values following PAMILO and NEI's distance (hereafter denoted by P_d) in the above two examples, we have 0.759 and 0.805 for one and two genes from each population, respectively. The value of P_d was computed as follows. Let d_{XY}, d_{YZ} and d_{XZ} be the average divergence times of genes from three pairs of different populations. These average divergence times are computed in the following way. For a pair of genes sampled from different populations, we can define the time at which there existed the most recent common ancestor. The time is averaged over all pairwise comparisons, providing d between a pair of populations. The consistency probability is then defined by

$$P_d = \text{Prob}(d_{XY} < d_{YZ} \text{ and } d_{XY} < d_{XZ}), \quad (18)$$

that is the probability of occurrence of gene trees in which the average genetic distance between X and Y is smaller than that between the other two combinations of populations. The difference between $P_d = 0.759$ and $P = 0.632$ for a three gene sample is that P_d includes the factor $1/3 \exp(-T_2) = 0.123$ that we have ignored as mentioned earlier. By the same token, the difference between the values of P and P_d for a six gene sample can be partly explained. However,

TABLE 3

Consistency probabilities between gene and population trees (simulation results with 10^5 repeats)

$t_1/(2N)$	$t_2/(2N)$	Sample size	P	P^*	P_d
0.05	0.05	1	0.049	0.384	0.384
		2	0.198	0.462	0.426
		5	0.602	0.727	0.433
		10	0.913	0.939	0.426
	0.5	1	0.379	0.575	0.575
		2	0.760	0.814	0.726
		5	0.993	0.994	0.806
		10	1.0	1.0	0.863
	5	1	0.994	0.998	0.998
		2	0.998	0.999	0.999
		5	1.0	1.0	1.0
		10	1.0	1.0	1.0
0.5	0.05	1	0.040	0.363	0.363
		2	0.118	0.401	0.370
		5	0.258	0.529	0.384
		10	0.347	0.563	0.386
	0.5	1	0.395	0.588	0.588
		2	0.640	0.757	0.697
		5	0.865	0.903	0.713
		10	0.921	0.952	0.716
	5	1	0.983	0.991	0.991
		2	0.999	0.999	0.999
		5	0.999	1.0	1.0
		10	1.0	1.0	1.0
5	0.05	1	0.045	0.367	0.367
		2	0.054	0.360	0.361
		5	0.048	0.344	0.344
		10	0.061	0.365	0.363
	0.5	1	0.390	0.589	0.589
		2	0.403	0.596	0.594
		5	0.378	0.593	0.592
		10	0.363	0.584	0.585
	5	1	0.992	0.993	0.993
		2	0.996	0.998	0.998
		5	0.997	0.997	0.997
		10	0.992	0.993	0.993

P, P^* and P_d are defined in (14), (19) and (18), respectively.

there is another factor that causes the difference. In PAMILO and NEI, there is an unresolvable case. It is "unresolvable" because their theory does not take account of the order and time of coalescences in the ancestral population. However, in actual data as well as simulations, we can always determine them and hence classify a gene tree into either a consistent or inconsistent class. Noting these differences and the probability of an unresolvable class ($R = 0.083$, see Table 2 in their paper), we can account for the difference between the values of P and P_d. Simulation results for the same sets of parameter values as in Table 2 are presented in Table 3.

DISCUSSION

We will first discuss some characteristics of the consistency probability P, defined based on the interspecific coalescence of genes (Table 2). As expected, if the time between the first and second population splitting (t_2) is long, the P value is close to 1 regardless of sample sizes. A sufficient condition for P to be close to 1 is that t_2 is not smaller than $10N$. In this case, a gene tree is almost surely consistent with the population tree and there is no need to increase sample sizes for a reliable estimate of the population tree. For smaller values of t_2, on the other hand, the P value strongly depends on sample sizes and t_1 (the divergence time between the two closest populations). When t_1 is small and the sample size is large, there remain many ancestors of genes sampled from populations X and Y at the time of their divergence. Then some of these ancestors will interspecifically coalesce during t_1 and $t_1 + t_2$, making the P value high. For instance, $P = 0.929$ in the case of $t_1 = t_2 = 0.1N$ and a sample of 10 genes from each population. Compared with the case of a sample of one gene from each population ($P = 0.049$), there is a dramatic increase in the P value by increasing sample size. However, when t_1 is large, the P value does not increase substantially. In this situation, there remains only one ancestor of genes from each population and the P value remains the same as that for a sample of one gene from each population. Thus, in general when t_1 does not much exceed N generations, a large sample can substantially increase the P value, making it possible to correctly infer the population tree.

The above conclusion is different from that in PAMILO and NEI (1988). There are two reasons for this discrepancy: actually they did not consider a sample of more than two genes from each population and used P_d defined by genetic distances between different populations. This restriction and criterion are connected to each other because the computation of genetic distances requires information on the topology of gene trees which are very difficult to analyze for arbitrary sample sizes. Recall that the genetic distance, defined by the average divergence time of genes sampled from different populations, is calculated based on all pairwise comparisons, in which all possible topologies of gene tree should be taken into account (*e.g.*, see TAKAHATA and NEI 1985). Because of this difficulty in the calculation of genetic distances, simulations were conducted and the results are given in Table 3 and Figure 2. The P_d value shows a rather weak dependence on sample size in a wide range of values of t_1 and t_2 even when the P value sharply increases as sample size increases. This insensitivity of P_d to sample size is due to its definition. As mentioned earlier, genetic distance tends to overshadow minor but phylogenetically useful information in a gene tree.

A large discrepancy between the values of P and P_d makes us suspicious about using the genetic distance in inferring a population tree. It is more promising to

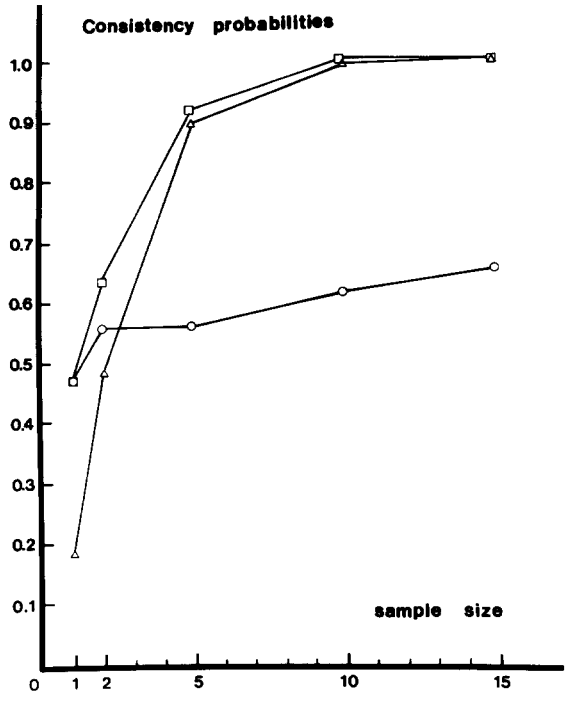

FIGURE 2.—Sample size dependence of the consistency probabilities obtained by simulations. The probabilities are defined in three different ways. Open triangles represent the probability P that at least one interspecific coalescence occurs during t_1 and $t_1 + t_2$ (see Equation 14 in text), while open circles represent P_d based on the average divergence times in all pairwise comparisons of genes from different populations. Open squares represent the probability P^* that the time on the first interspecific coalescence of genes from X and Y is shorter than that from X and Z and from Y and Z. Here $t_1 = 0.2N$ and $t_2 = 0.4N$.

use interspecific coalescences instead. For this end, one problem arising from the fact that a population tree is actually unknown must be solved. If the divergence times t_1 and t_2 of populations are unknown, it is uncertain whether the first interspecific coalescence (point A in Figure 1) occurred during the time between t_1 and $t_1 + t_2$. A gene tree constructed from nucleotide differences does not have such a time ruler as depicted in the ordinate in Figure 1. However, it does tell us the order of the first interspecific coalescences from different pairs of populations. This is information we can use in inferring the population tree. Let τ_A be the first interspecific coalescence time for genes from population X and Y, and τ_B that for population Z and X (or Y) ($\tau_A \geqq t_1$ and $\tau_B \geqq t_1 + t_2$). We are interested in the probability of $\tau_A < \tau_B$ in a gene tree and use it to infer the population tree:

$$\begin{aligned} P^* &= \text{Prob}(\tau_A < \tau_B) \\ &= \text{Prob}(\tau_A < t_1 + t_2 \leqq \tau_B) \\ &\quad + \text{Prob}(t_1 + t_2 \leqq \tau_A < \tau_B) \\ &= P + Q. \end{aligned} \tag{19}$$

The first term of the right hand side in (19) is the probability that we formulated in (14), and the second term corresponds to the event that τ_A is smaller than τ_B by mere chance. For a sample of one gene from each population, $P = 1 - \exp(-t_2/(2N))$ and $Q = 1/3 \exp(-t_2/(2N))$ so that $P^* = 1 - (2/3)\exp(-t_2/(2N))$ as derived in NEI (1987). The P^* value for arbitrary samples was obtained by simulation (Table 3 and Figure 2). By definition $P^* \geqq P$, and it is clear that P^* has the same dependence on sample sizes as P, although a large difference between P* and P is expected when most interspecific coalescences occur before $t_1 + t_2$.

The values of $T_1 = t_1/(2N)$ and $T_2 = t_2/(2N)$ in Figure 2 were chosen to mimic the population tree of three human races (NEI and ROYCHOUDHURY 1982; PAMILO and NEI 1988). It is remarkable in this figure that $P^* = 0.9$ is attained for a sample of five genes from each population whereas P_d is about 0.6 and stays around the same value for further increases in sample size. NEI (1985, 1987) presented a phylogenetic tree of 10 mtDNAs from each of Caucasoid (X), Mongoloid (Y), and Negroid (Z), and suggested using average numbers of nucleotide differences between different populations for finding the order of population splitting. The average nucleotide differences were then estimated as $d_{XY} = 0.308\%$, $d_{YZ} = 0.416\%$ and $d_{XZ} = 0.379\%$ (see Table III in NEI 1985). If we assume that the average number of nucleotide differences is in proportion to the average divergence time of genes (genetic distance), these figures in fact support the closer relationship between Caucasoid and Mongoloid, but the P_d value for this example is only 0.6 (Figure 2). However, if we take a close look at the gene tree given in Figure 10.5 in NEI (1987) with respect to interspecific coalescences, the condition for (19) is satisfied so that we can assert the same phylogenetic relationship among three human races with 90% confidence. To show this large difference between P^* and P_d, Figure 3 was drawn. This is a gene tree generated on a computer under the same condition as in Figure 2 with a sample size of 10 for each population. It demonstrates a case where $d_{XY} > d_{YZ}$ or d_{XZ} but $\tau_A < \tau_B$, and is a typical pattern of gene tree expected under neutrality. We thus conclude that a population tree can be inferred more reliably by using interspecific coalescences than by using genetic distance.

We have assumed that populations are in a stationary state. If a population undergoes a bottleneck, genes at a locus drawn from a current population may have been derived from a common ancestor that existed during such a contracted phase of population size. Bottlenecks disrupt the stationarity of population and hence may change a gene tree in a significant way. Here we assert only two things in order for the effects of bottleneck to be manifest in gene genealogy. First, it must occur relatively recently. If the occur-

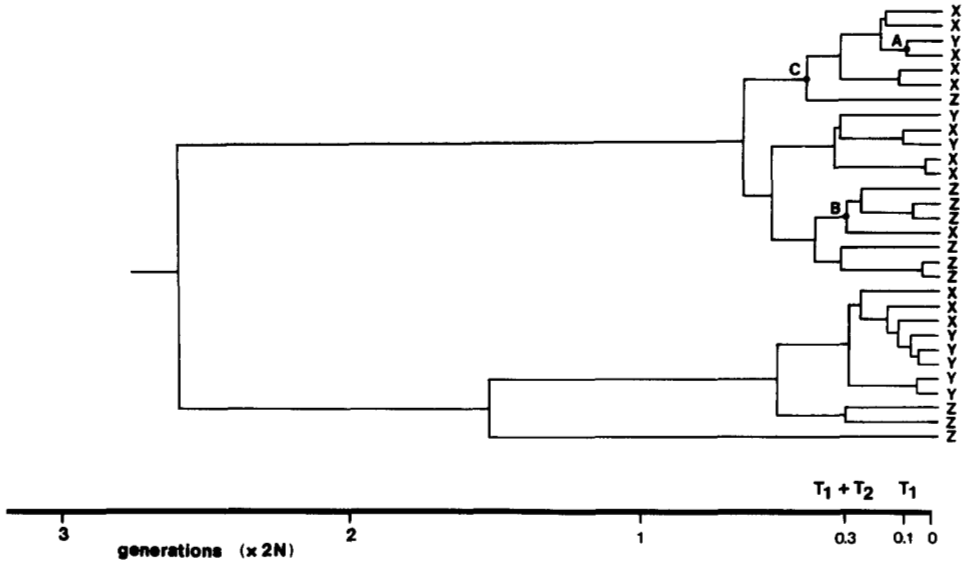

FIGURE 3.—Gene tree generated on a computer. Ten genes were sampled from each of three populations, and $t_1 = 0.2N$ and $t_2 = 0.4N$ were assumed. In this simulation, $d_{XY} = 6.0N$, $d_{YZ} = 5.7N$ and $d_{XZ} = 6.5N$ were observed. Thus the closer relationship between Y and Z is indicated in terms of the average divergence times of genes, which is inconsistent with the population tree. By contrast, this gene tree becomes consistent in terms of interspecific coalescences (Note points A, B and C).

rence is sufficiently long time ago compared with the current population size, most genes were derived from a common ancestor which existed after the bottleneck and thus the genealogy is independent of such a remote event. Secondly, even if a bottleneck occurred recently, the effect can be seen only when the duration time is long enough compared with the reduced population size. For instance, if the reduced population size is 100, then the required duration time is also at least about 100 generations (APPENDIX). Clearly, the effect of bottlenecks on the consistency probability depends on when and how strongly they have occurred in the history of populations.

Now we ask a question on sampling strategy: Is it necessary to examine many independent loci or sufficient to examine many genes at a single locus? The answer depends on whether we use genetic distance or interspecific coalescence, as well as on the values of t_1 and t_2. If we use genetic distance, we come to the same conclusion as PAMILO and NEI (1988): to obtain a reliable population tree, one must study many genes which have evolved independently of each other. As demonstrated above, this is largely due to the poor performance of the metric used. If on the other hand we use interspecific coalescence, we come to a different conclusion. To argue this point quantitatively, assume that n independent loci were examined and consider the probability P_T that at least one of the loci shows consistency between gene and population trees, where P is used for the consistency probability. The reason for using P is that if at least one of the loci shows interspecific coalescence between t_1 and $t_1 + t_2$, the order of population splitting becomes certain. In this sense, we do not follow the majority rule as in SAITOU and NEI (1986) in which the correct population tree is regarded as the one represented by the largest number of loci. Then we have a simple formula

$$P_T = 1 - (1 - P)^n. \qquad (20)$$

For a sample of one gene from each population, P_T becomes $1 - \exp(-nT_2)$ so that for P_T to be larger than 0.95, n must be larger than $3/T_2$. Thus $n > 60$ for $T_2 = 0.05$ and $n > 15$ for $T_2 = 0.2$. Although these numbers may not be too unrealistic, recall the case of $T_2 = 0.2$ in Figure 2 where 10 genes at a locus can confirm the closer relationship between X and Y when we find $\tau_A < \tau_B$ in the gene tree ($P^* \cong P$ in this situation). Thus large samples at a locus can provide very useful information on a population tree. Moreover, large samples allow us to estimate population sizes which are indispensable parameters in any theory. The only situation in which the present method does not work is where t_1 is large but t_2 is small relative to population sizes. It is, however, the case where the three populations practically diverged around the same time and sampling several independent loci does not resolve the problem either.

We have studied the consistency probability between gene and population trees, assuming that there are no stochastic errors in a gene tree other than those caused by random drift. In practice, any gene tree constructed from nucleotide differences involves stochastic errors owing to mutations. It is therefore interesting to see whether our conclusion remains true when mutational errors are incorporated. I conducted a simulation in which mutations following Poisson processes are superimposed on a gene tree and examined P^* and P_d in terms of nucleotide differences (Figure 4). It was assumed that $2Nv = 10$ where v is the mutation rate per gene (or linked DNA segment) per generation. Since $T_1 = 0.1$ and $T_2 = 0.2$ were used as in Figure 2, the average number of mutations that accumulate per gene during these times amounted to 1 and 2, respectively. These numbers are indeed very small relative to the extent of intrapopulational variation, yet it is clear that the sample

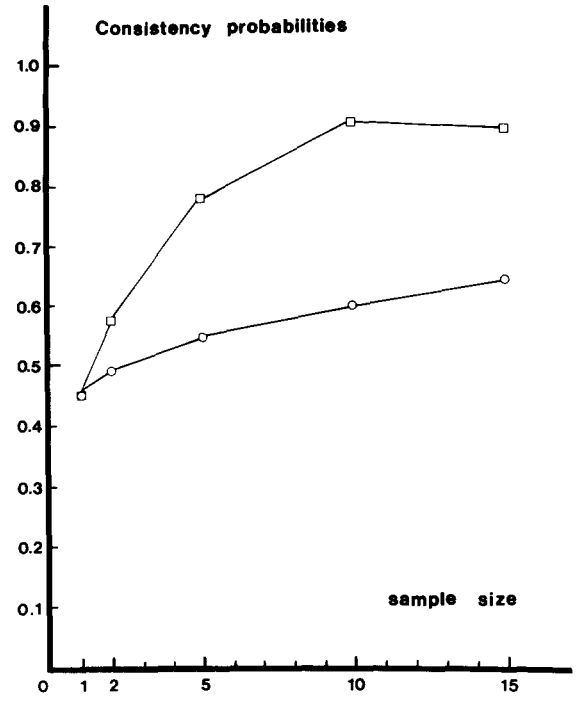

FIGURE 4.—Consistency probabilities P^* (open squares) and P_d (open circles) when they were defined in terms of nucleotide differences. As in Figures 2 and 3, $t_1 = 0.2N$ and $t_2 = 0.4N$ but the mutation rate v is assumed to be $5/N$ per gene per generation.

size dependence of P^* and P_d is essentially the same as before. If we reduce the value of $2Nv$, however, the P^* value diminishes, implying that a large number of linked nucleotide sites must be examined (SAITOU and NEI 1986). In the case of human mtDNA, the average nucleotide difference per site within the population is about 0.36% so that the number of nucleotide sites examined must be about 2,800 for $2Nv$ to be 10. Together with this requirement, the present theory will hopefully help improve experimental designs for the problem treated in this paper.

I thank M. NEI, N. SAITOU and two anonymous reviewers for their comments on an early version of this paper. This work is supported in part by grants from the Ministry of Education, Science and Culture in Japan and from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

DONNELLY, P., 1984 The transient behaviour of the Moran model in population genetics. Math. Proc. Camb. Phil. Soc. **95:** 349–358.

KIMURA, M., 1986 Evolutionary rate at the molecular level. Nature **217:** 624–626.

KINGMAN, J. F. C., 1982 The coalescent. Stochastic Process. Appl. **13:** 235–248.

NEI, M., 1985 Human evolution at the molecular level, pp. 41–64 in *Population Genetics and Molecular Evolution*, edited by T. OHTA and K. AOKI. Japan Scientific Societies Press, Tokyo.

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

NEI, M., and A. K. ROYCHOUDHURY, 1982 Genetic relationship and evolution of human races. Evol. Biol. **14:** 1–59.

NEIGEL, J. E., and A. C. AVISE 1986 Phylogenetic relationships of mitochondrial DNA under various models of speciation, pp. 515–534 in *Evolutionary Processes and Theory*, edited by S. KARLIN and E. NEVO. Academic Press, New York.

PAMILO, P., and M. NEI, 1988 Relationships between gene trees and species trees. Mol. Biol. Evol. **5:** 568–583.

SAITOU, N., and M. NEI, 1986 The number of nucleotides required to determine the branching order of three species with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. **24:** 189–204.

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

TAKAHATA, N., AND M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

TAVARÉ, S., 1984 Lines-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

WATTERSON, G. A., 1984a Allele frequencies after a bottleneck. Theor. Popul. Biol. **26:** 387–407.

WATTERSON, G. A., 1984b Lines of descent and the coalescent. Theor. Popul. Biol. **26:** 77–92.

Communicating editor: E. THOMPSON

APPENDIX

We assume that the size of a diploid population changes abruptly t_1 and $t_1 + t_2$ generations ago, with the population size being N_1 for $0 \leqq t \leqq t_1$, N_2 for $t_1 < t \leqq t_1 + t_2$ and N_3 for $t > t_1 + t_2$. The generation of the population at $t_1 + t_2$ consists of $2N_2$ genes chosen at random with replacement from the $2N_3$ genes from the previous generation. Similarly, the generation at t_1 consists of $2N_1$ genes chosen at random with replacement from the $2N_2$ genes of the previous generation. Let A_t be the number of distinct ancestors of sampled genes t generations ago. Our aim is to evaluate the probability of $A_t = j$ at $t = t_1 + t_2$, given $A_0 = i$ $(1 \leqq j \leqq i)$. WATTERSON (1984a) considered a similar but more complicated problem that arises when effects of mutations are incorporated in the genealogical process.

We denote by $g_{ij}(t, N)$ the probability of $(A_t = j \mid A_0 = i)$ in a population of size N. For a stationary population, it is given in Tavaré (1984), DONNELLY (1984), WATTERSON (1984b), and TAKAHATA and NEI (1985). The formula of $g_{ij}(t, N)$ has an invariance property, which is that for an arbitrary constant c,

$$g_{ij}(t, N) = g_{ij}(ct, cN). \tag{A1}$$

In words, (A1) implies that a gene tree in a c times larger population is exactly c times magnified compared with that in a population of size N. Another important property of A_t is Markovian, that is for any times r and s,

$$g_{ij}(r + s, N) = \sum_{k=j}^{i} g_{ik}(r, N) g_{kj}(s, N). \tag{A2}$$

Equating A2 holds true whether or not r $(0 \leqq r \leqq r + s)$ is the time of coalescence, and it is due to the fact that the time between two successive coalescences is exponentially distributed (KINGMAN 1982). Using (A2), we can readily express the prob-

ability $\hat{g}_{ij}(t) = \text{Prob}(A_t = j \mid A_0 = i)$ for the present nonstationary population (indicated by a caret over g_{ij}):

$$\hat{g}_{ij}(t) = g_{ij}(t, N_1), \quad 0 < t \leqq t_1$$

$$= \sum_{k=j}^{i} g_{ik}(t_1, N_1) g_{kj}(t - t_1, N_2),$$

$$t_1 < t \leqq t_1 + t_2 \tag{A3}$$

$$= \sum_{k=m}^{i} \sum_{m=j}^{k} g_{jk}(t_1, N_1) g_{km}(t_2, N_2) g_{mj}(t - t_1 - t_2, N_3),$$

$$t > t_1 + t_2.$$

Of particular interest here is $\hat{g}_{ij}(t)$ at $t = t_1 + t_2$, which is given as

$$\begin{aligned} \hat{g}_{ij}(t_1 + t_2) &= \sum_{k=j}^{i} g_{ik}(t_1, N_1) g_{kj}(t_2, N_2) \\ &= \sum_{k=j}^{i} g_{ik}(t_1, N_1) g_{kj}(ct_2, N_1) \\ &= g_{ij}(t_1 + ct_2, N_1). \end{aligned} \tag{A4}$$

In the above, $c = N_1/N_2$ and we have used (A1) and (A2). The model of bottlenecks assumes that $N_1 > N_2$ so that the genealogical process A_t speeds up c times during the bottleneck phase. Although the general formula of $g_{ij}(t, N)$ is rather complicated, it is useful to record the probability of no coalescence during the bottleneck phase, which is given by

$$g_{kk}(t_2, N_2) = \exp\left\{\frac{-k(k-1)t_2}{4N_2}\right\}. \tag{A5}$$

Thus the strength of the bottleneck can be evaluated by whether or not $k(k-1)t_2 \gg N_2$. Clearly the smaller N_2 and the longer t_2, the more likely the condition is satisfied.