Letters to the Editor

Gene Trees, Species Trees and the Segregation of Ancestral Alleles

WU (1991) has recently calculated a probability, denoted by $1 - P(T)$, which he describes as the probability of obtaining incorrect phylogenetic information for three species due to segregation of ancient polymorphism. To obtain $P(T)$, he considers three species with the relationship shown in Figure 1 and supposes that one sequence is obtained from each species. I will designate the three sequences by $s1$, $s2$ and $s3$. $P(T)$, as calculated by WU, is actually the probability that either α or γ occurs, given that one of the events, α, γ, β_1 or β_2 occurs. These four events are defined as follows:

α: The event that $s1$ and $s2$ are descendants of allele A_i and $s3$ is a descendant of $A_{\sim i}$, where A_i is an allele that was segregating in the ancestral population at node 1 of the species tree in Figure 1, and where $A_{\sim i}$ is any other allele.

γ: The event that $s1$ and $s2$ are both descendants of a mutant that arose between node 1 and node 2 of the species tree.

β_1: The event that $s1$ and $s3$ are descendants of allele A_i and $s2$ is a descendant of $A_{\sim i}$.

β_2: The event that $s2$ and $s3$ are descendants of allele A_i and $s1$ is a descendent of $A_{\sim i}$.

Note that α and γ are not mutually exclusive, so it is possible for both α and γ to occur in the history of the sequences $s1$, $s2$ and $s3$. However, if the mutation rate is small, then the probability that both α and γ occur is negligible, in which case, $P(T)$, which WU describes as the probability of correct phylogenetic information is

$$P(T) \approx \frac{P(\alpha) + P(\gamma)}{P(\alpha) + P(\gamma) + P(\beta_1) + P(\beta_2)}.$$

WU calculates each of the probabilities on the right hand side under a Wright-Fisher neutral model and finds, for small mutation rates that

$$P(T) \approx \frac{1 + T}{1 + T + 2_e^{-T}}, \quad (1)$$

where T is measured in units of $2N$ generations, N being the diploid population size. Equation 1 is the main result of the first part of WU's paper. I claim that this conditional probability is not appropriate for the interpretation of any experimental observations because the events, α, γ, β_1 or β_2, considered by WU, do not correspond in an appropriate way to any observable pattern in data. It will be shown below, for example, that the events α, β_1 and β_2, can produce data that is indistinguishable from data resulting from events other than α, γ, β_1 or β_2. Therefore, knowing the conditional probability, $P(T)$, is not useful for assessing the probability of particular patterns in data.

First, it is important to note that there are two very distinct situations, which must be distinguished in analyzing this problem: *situation 1*, in which one does not distinguish derived and ancestral states of the sampled genes, and *situation 2*, in which one does distinguish derived and ancestral states of the sampled genes. Clearly, in situation 2 more patterns are discernible than in situation 1, and so these two situations must be treated separately.

If one has sequences from only three species, one is typically in situation 1, being unable to determine the ancestral state of the genes. Thus, for example, if $s1$ and $s3$ are alike in missing a contiguous set of 20 bp relative to $s2$, one cannot tell if, on the one hand, there was a deletion which occurred in a common ancestor of $s1$ and $s3$, or on the other hand, an insertion occurred in the lineage leading to $s2$. (I am ignoring the possibility of the same insertion or deletion occurring twice or an exact reversal of an earlier insertion or deletion.) If it was an insertion that occurred in the lineage leading to $s2$, this insertion could have occurred quite recently, at the time indicated by the filled square in Figure 1, or much farther in the past, at the open square. Note that in situation 1, with low mutation rate, there are only three likely patterns other than all three sequences alike, namely, $s1$ and $s2$ alike with $s3$ different, $s1$ and $s3$ alike with $s2$ different, and $s2$ and $s3$ alike with $s1$ different. These three patterns will be designated (12)3, (13)2 and (23)1, respectively. Other patterns require more than one mutation which are very unlikely with low mutation rates.

Can an investigator in situation 1 use $P(T)$ to assess the probability of pattern (12)3, or one of the other two patterns, under the assumption that the species tree is the one shown in Figure 1? The answer is no. When α or γ occurs, then pattern (12)3 will necessarily result. An example is shown in Figure 1, where if a mutation occurs at the open circle, pattern (12)3 will be produced. However, the pattern (12)3 can arise without α or γ occurring. An example is when a mutation occurs at the point indicated by the closed circle, in which case none of the events, α, γ, β_1 or β_2 has occurred. An investigator in situation 1, cannot distinguish between a mutation at the open circle and a mutation at the closed circle. Therefore, to interpret

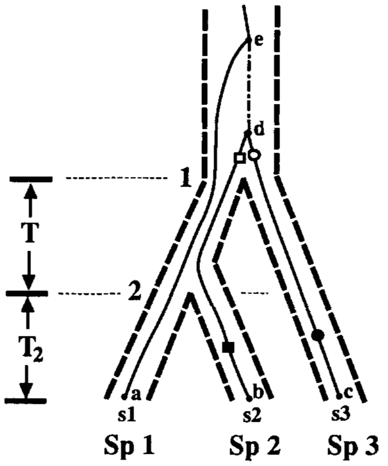

FIGURE 1.—A species tree with an example gene tree. The species tree is shown by the bold dashed lines, with the speciation events indicated by the 1 and 2. The sampled genes from the species are indicated as $s1$ and $s2$ and $s3$. The gene tree is drawn with smaller solid and "dot-dash" lines. Mutations that occur on the dot-dash line result in shared derived states in two sequences. The squares and circles represent mutations discussed in the text.

data in situation 1, one needs the probability of (12)3 or perhaps the probability of (12)3 given that either (12)3, (13)2 or (23)1 has occurred, rather than the probability of α or γ or the conditional probability, $P(T)$. In order to compare these probabilities, I now calculate the probability of (12)3 and the other two patterns under a Wright-Fisher neutral infinite-allele model with low mutation rate.

Under the infinite-allele model, the probability of pattern (12)3 for the gene tree of Figure 1 is the probability of at least one mutation on the branch cd and no mutations on the rest of the gene tree. (In the following, a branch of a gene tree will be referred to by the pair of letters which label the ends of the branch in Figure 1. The length of branch ij, measured in units of $2N$ generations will be denoted by $L(ij)$.) Thus, the probability of pattern (12)3 given the gene tree in Figure 1 is

$$P((12)3\,|\,\text{gene tree 1}) = (1 - e^{-(M/2)L(cd)}) \cdot e^{-(M/2)(L(ae) + L(de) + L(bd))} \approx \frac{M}{2} L(cd), \tag{2}$$

where M is $4Nu$, and u is the neutral mutation rate for the genetic region sequenced. The approximation is for M small. For other gene trees and other patterns, similar expressions hold. For example, the probability of (23)1 given the gene tree of Figure 1, is approximately $(M/2)$ times the sum of the lengths of branches ae and de. The unconditional probabilities of the patterns can be obtained by taking the expectation over all possible gene trees and branch lengths. All branch lengths are distributed approximately exponentially, assuming N is large (KINGMAN 1982), which makes the calculation of the unconditional probabilities straightforward. For M sufficiently small, the unconditional probabilities of the three patterns are:

$$P((23)1) = P((13)2) \approx \frac{M}{2}(T_2 + 1) \tag{3}$$

and

$$P((12)3) \approx \frac{M}{2}(T_2 + 2T + 1). \tag{4}$$

From (3) and (4), the probability of (12)3 conditional on one of the three patterns, (12)3, (23)1, (13)2 occurring is given approximately by

$$P((12)3 \mid (12)3, (13)2 \text{ or } (23)1) \approx \frac{T_2 + 2T + 1}{3T_2 + 3 + 2T}. \tag{5}$$

These equations show, what is intuitively clear, that the probability of pattern (12)3 depends on T_2 as well as T. It is also intuitively clear that when T_2 is large compared to T, then all three patterns must be about equally likely, even if T is much larger than $2N$ generations. Equation 5 bears this out, since the right hand side is approximately ⅓ when T_2 is large compared to T. These results are very different from WU's results. Equations 2–5 rely on the assumption that M, and MT and MT_2 are all small, which may be appropriate for large insertion/deletion events. However, for nucleotide substitutions it is probably necessary to consider a finite-allele model and incorporate the possibility of multiple hits.

Summarizing, in situation 1, the conditional distribution of observable patterns is very different from the conditional probability obtained by WU. WU's conditional probability would not be appropriate for interpreting patterns observed in situation 1.

Now I consider situation 2. If one has data from three species plus an outgroup species, and if one can assume that mutations are unique and irreversible, then one is in situation 2, *i.e.*, the ancestral state of genes can be determined. In this case there are 6 possible patterns produced by a single mutation on the gene tree. Three of the possible patterns have two of the sequences sharing a derived state and the third sequence retaining the ancestral state. The other three patterns have one sequence with a derived state and the other two retain the ancestral state. Subscripts a and d will be used to designate ancestral and derived states of the sequences. For example, $(13)_d2_a$ will be used to designate that $s1$ and $s3$ share a derived state with $s2$ retaining the ancestral state. $(13)_a2_d$ will denote the case where $s1$ and $s3$ retain the ancestral state and $s2$ is derived.

Can an investigator in situation 2, use Equation 1, to assess the probability of pattern $(12)_d3_a$, or any of the other five possible patterns, under the assumption that the species tree is the one shown in Figure 1? Again the answer is no. The probability of α plus the probability of γ does not equal the probability of

$(12)_d3_a$. While the pattern $(12)_d3_a$ implies that α or γ has occurred, and γ necessarily results in $(12)_d3_a$, it is not true that α necessarily produces pattern $(12)_d3_a$. Therefore the probability of α plus the probability of γ is greater than the probability of $(12)_d3_a$. To see that α does not necessarily produce pattern $(12)_d3_a$, consider the gene tree of Figure 1. If a mutation occurred at the open circle, then α has occurred, but the resulting pattern is $(12)_a3_d$ not $(12)_d3_a$. Notice that the pattern $(12)_a3_d$, which can be produced by α, can also be produced by events other than α, γ, β_1 or β_2, for example, when a mutation occurs at the solid circle of the gene tree in Figure 1.

Notice that when two of the three sequences share a derived state, the topology of the gene genealogy is unambiguously determined. If the species tree is the tree shown in Figure 1, and one observes $(13)_d2_a$ or $(23)_d1_a$ then it is necessarily true that the gene genealogy is incongruent with the species tree, and an ancestral polymorphism in the population at node 1 is necessarily involved. So, if an investigator observes shared derived states in two of the sequences, then a useful quantity for assessing the probability that the gene tree (which is now established) has the same topology as the species tree is the following conditional probability:

$$R = \frac{P((12)_d3_a)}{P((23)_d1_a) + P((13)_d2_a) + P((12)_d3_a)}.$$

One can calculate the conditional probability, R, under the same infinite-allele model considered above for situation 1, however it is more informative to analyze an infinite-site model. In this model, only one mutation can occur at any particular site, but more than one mutation can occur in the whole region sequenced. The mutation rate at individual sites is infinitesimal, but M, the mutation parameter for the entire region sequenced, is not necessarily small. For M small, the results will converge to what would be obtained under an infinite-allele model with small mutation parameter. For nucleotide substitutions the mutation rate per site may not be sufficiently small, in which case one should consider a many-site model with each site being modeled as a four-allele locus. However, for large insertion/deletion events the infinite-site model may be sufficiently accurate. We are, in essence, assuming that an insertion/deletion at a particular site only happens once on the gene genealogy and that once an insertion/deletion occurs, subsequent insertion/deletions do not completely obscure the earlier event.

For the infinite-site model, I will use $(23)_d1_a$ to denote that case where one *or more* sites in the region sequenced show the pattern $(23)_d1_a$. Similarly, R denotes the probability that $s2$ and $s3$ share a derived state and $s1$ retains the ancestral state, at one *or more* sites, conditional on any pair of sequences sharing a derived state at one or more sites. R, defined in this way, may be a useful quantity for assessing the likelihood of various species trees when one has sequences which exhibit shared derived states, and for this reason I will obtain an expression for it below. However, one could presumably make more informed assessments of the likelihood of different species trees by considering the number of sites in the data which show pattern $(23)_d1_a$, as well as the number of sites that show each of the other possible patterns. This problem will not be pursued here.

To obtain R, I begin by calculating the probability of $(23)_d1_a$. This pattern can only arise if the gene tree is like the gene tree shown in Figure 1 and if one or more mutations occur on branch *de* of that tree. That is, the most recent common ancestor of $s1$ and $s2$ must have occurred before node 1. This occurs with probability e^{-T}. In addition, the topology of the gene tree must be such that $s2$ and $s3$ are the most recently diverged pair of sequences. This occurs with probability ⅓, conditional on the most recent common ancestor of $s1$ and $s2$ occurring before node 1. Given these first two conditions, the probability of at least one mutation on the branch *de* is $M/(2 + M)$, which follows from the fact that the duration of branch *de* is exponentially distributed with mean 1, in units of $2N$ generations. Thus,

$$P((23)_d1_a) = P((13)_d2_a) = \left(\frac{1}{3}\right)\left(\frac{M}{2+M}\right) e^{-T}. \quad (6)$$

The probability of $(12)_d3_a$ is slightly more complicated since the mutations which lead to the shared derived state of $s1$ and $s2$ can occur either before or after node 1. Taking these possibilities into account, one finds

$$P((12)_d3_a) = 1 - e^{-T} + \left(\frac{M}{2+M}\right)\frac{e^{-T}}{3} + \frac{4(e^{-T} - e^{-MT/2})}{(2-M)(2+M)}. \quad (7)$$

Using (6) and (7), the conditional probability, R, now written as $R(T,M)$, to indicate its dependence on T and M, is found to be:

$$R(T,M) = 1 - \frac{2e^{-T}}{3} \cdot \left\{\frac{M(2-M)}{(2+M)(2-M) - 4e^{-MT/2} + 2Me^{-T}}\right\}. \quad (8)$$

The limit of $R(T,M)$ as M tends to zero, gives the small M approximation:

$$R(T,0+) \approx 1 - \frac{2e^{-T}}{3}\left\{\frac{1}{T + e^{-T}}\right\}. \quad (9)$$

Equation 9 can also be obtained by WU's approach

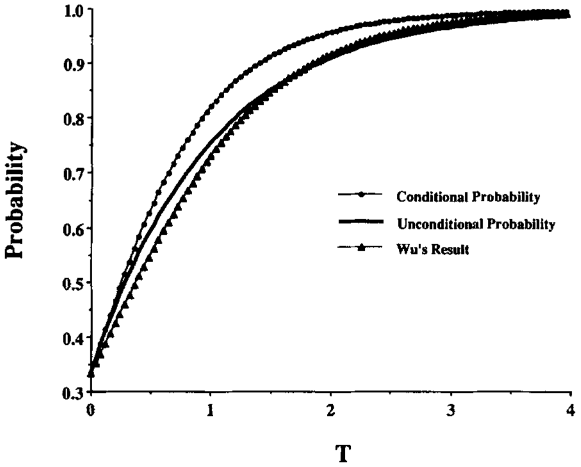

FIGURE 2.—The probability that the gene tree is congruent with the species tree, as a function of T, the time between node 1 and node 2 in Figure 1. The unconditional probability is from Equation 10, the conditional probability from Equation 9, and WU's result is Equation 1.

of considering alleles at particular frequencies at node 1. To do this, it is helpful to consider α', β_1' and β_2' instead of α, β_1 and β_2. α' is the event that $s1$ and $s2$ are descendants of the same *derived* allele present at node 1 and $s3$ is a descendant of an ancestral allele. Similarly, β_1' and β_2' are defined like β_1 and β_2 except that the ancestral and derived states are specified as for α'. The conditional probability of α' or γ given α', γ, β_1' or β_2' is the same as my R. WU's Equations 1 and 2 can be modified to calculate the probability of β_1' and α', when mutation rates are low, by multiplying the right hand side of these equations by $(1 - p)$. With low mutation rates, $1 - p$ is the probability that the allele at frequency p is a derived allele. This follows from results of WATTERSON and GUESS (1977) on the ages of alleles. Following through with WU's analysis, but with the modified versions of his Equations 1 and 2, leads to Equation 9, above.

Equations 8 and 9, which give the probability of gene tree and species tree congruence conditional on two of the three sequences sharing a derived state, should be compared to the unconditional probability of gene tree and species tree congruence

$$P_{(\text{congruent gene tree})} = 1 - \frac{2e^{-T}}{3}, \qquad (10)$$

(HUDSON 1983; NEI 1986). Equations 1, 9 and 10 are plotted in Figure 2. Note that for M large there will always be at least one mutation producing a shared derived state in the sampled sequences, so the conditional probability, $R(T,M)$, should approach the unconditional probability given by (10). This is indeed the case, since the expression in curly brackets on the right hand side of (8) approaches one as M goes to infinity. Notice that the expression in curly brackets on the right hand side of (9) is always less than one for $T > 0$, and thus, for M small, the conditional probability of congruence is always greater than the unconditional probability of congruence (see Figure 2). Using (9), it is found that, conditional on finding two of the sequences with a share derived state, the probability of the gene tree and the species tree being congruent is greater than 0.95 for $T > 1.9$. This contrasts with WU's result that T must be greater than 2.4 for 95% confidence in congruence of gene tree and species tree. From equation (10), the unconditional probability that the gene tree is congruent with the species tree is greater than 0.95 for $T > 2.6$.

Can we consider WU's result for small M to be an approximation for equation (9)? In Figure 2 one can compare WU's result with the conditional probability and the unconditional probability. This figure shows that the unconditional probability of congruence is closer to the conditional probability than is WU's result for a substantial range of T values, suggesting that Wu's result should not be considered as an approximation to this conditional probability.

In conclusion, WU's result is not appropriate for the interpretation of data in either situation 1 or situation 2. If mutation rates per site are sufficiently small, then Equation 5 should be used in situation 1 and Equation 8 or 9 in situation 2. It should be noted that my criticisms apply primarily to the first half of WU's paper which deals with a single locus. The second half of the paper which considers multiple loci is essentially independent of the first part of the paper. However, the maximum likelihood estimates of T which appear in the second half of the paper do depend on the first half of the paper and are therefore incorrect.

I am grateful to JOE FELSENSTEIN who made several useful suggestions, not all of which were incorporated in this letter. This work was supported by National Institute of Health grant GM42447.

RICHARD R. HUDSON
Department of Ecology and Evolutionary Biology
University of California
Irvine, California 92717

LITERATURE CITED

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

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

NEI, M., 1986 Stochastic errors in DNA evolution and molecular phylogeny, in *Evolutionary Perspectives and the New Genetics.* Alan R. Liss, New York.

WATTERSON, G. A., and GUESS, H. A., 1977 Is the most frequent allele the oldest? Theor. Popul. Biol. **11:** 141–160.

WU, C. I., 1991 Inferences of species phylogeny in relation to segregation of ancient polymorphism. Genetics **127:** 429–435.