Phylogenetic Analysis Under Reticulate Evolution

Shizhong Xu
Department of Botany and Plant Sciences, University of California at Riverside

The usual assumption that species have evolved from a common ancestor by a simple branching process—where each branch is genetically isolated—has been challenged by the observation of frequent hybridization between species in natural populations. In fact, most plant species are thought to have hybrid origins. This reticulate pattern of species evolution has posed problems in the definition of speciation and in phylogenetic reconstruction, especially when molecular data are used. As a result, hybridization has been largely treated as an evolutionary accident or statistical error in phylogenetic analysis. In this paper, I explicitly incorporate hybridization as an evolutionary occurrence and then conduct phylogenetic reconstruction.

I first examine the reticulate evolution under a pure drift model, and then extend the theory to fit a mutation model. A least-squares method is developed for reconstructing a reticulate phylogeny using gene frequency data. The efficacy of the method under the pure drift model is verified via Monte Carlo simulations.

Introduction

Species are often assumed to have evolved from a common ancestor by a complete process of branching, followed by complete genetic isolation (Cavalli-Sforza and Edwards 1967). A period of hybridization shortly after speciation is not thought to weaken this assumption provided the period is short relative to the periods between successive speciation events. However, species may hybridize long after speciation (Ritland and Eckenwalder 1992), which may pose problems in phylogenetic reconstruction, especially when molecular data are used (Spence 1990). In fact, many plant species are thought to have hybrid origins. The discovery of cytoplasmic introgression and the nonconcordance between the rDNA and cpDNA phylogenies of several plant groups is reflective of past hybridization and subsequent introgression (Systsma 1990; Rieseberg 1991; Rieseberg, Whitton, and Linder 1996).

McDade (1990, 1992) carried out artificial hybridization experiments among nine Central American species of the *Aphelandra pucherrima* complex and obtained 17 hybrid populations. She did extensive cladistic analysis, concluding that a hybrid will be placed by phylogenetic analysis as a basal lineage to the clade that includes its most derived parent. The phylogeny of 24 inbred strains of mice inferred by Atchley and Fitch (1991, 1993) includes several strains with hybrid origins. Atchley and Fitch found that a hybrid strain is always placed close to one of its parental strains. However, analytical tools that we use at present cannot generate reticulate diagrams that accurately depict a hybrid history (McDade 1995). If an analysis includes hybrids, no matter where the hybrids are placed, a cladistic method produces only divergently branching phylogenetic patterns and thus can never give the correct phylogeny. When traditional cladistic methods are applied anyway, they can give confusing and conflicting results. One typical result is that a large set of very different phylogenies will appear to be equally good (Hein 1990).

After carefully examining the cladistic behavior of hybrids, Funk (1985) provided some guidelines and methods for identifying possible hybrids in a cladistic study. Morefield (in Rieseberg and Morefield 1995) has developed a computer program, RETICLAD, that can identify hybrids based on the expectation that they will combine the characters of their parents. RETICLAD basically represents a quantification of Funk’s (1985) approach to recognizing hybrids. Rieseberg and Ellstrand (1993) provide a number of examples in which the program seems to work well. However, the RETICLAD program only tests reticulate events between terminal branches; hybridization between internal branches cannot be analyzed (Rieseberg and Morefield 1995).

In this paper, I focus on some theoretical aspects of reticulate evolution and develop a method using a simple reticulate phylogeny of four (or five) taxa (including one hybrid) as an example. I demonstrate the phylogenetic method under the pure drift model, and then extend the method to fit a mutation model. A least-squares method is developed to reconstruct the reticulate phylogeny using gene frequency data. The efficacy of the method under the drift model is verified via Monte Carlo simulation experiments. Finally, the joint effect of genetic drift and mutation on analysis of reticulate phylogeny is discussed.

The Pure Drift Model

In the absence of selection, mutation, and migration, the genetic composition of a finite population will change overtime in a random fashion. The heterozygosity within a single population, however, is a monotonic decreasing function of the time since the population is isolated from the base population. We have demonstrated that heterozygosities can be used to infer phylogenetic relationships among populations (Xu, Atchley, and Fitch 1994).

Reduction of Heterozygosity

In contrast to mutation, genetic drift consumes heterozygosity as the population evolves. Consider a finite
population with an effective population size of \( N_e \) that was isolated from an infinite random mating population \( t \) generations ago. The heterozygosity of the current population, denoted by \( H_t \), is

\[
H_t = H_0 \gamma^t, \tag{1}
\]

where \( H_0 \) is the heterozygosity of the base population and \( \gamma = 1 - 1/(2N_e) \) is the rate of heterozygosity reduction per generation. Equation (1) implies that if the loss of heterozygosity is known, one can infer the time since the population was isolated from the base population.

Population Divergence

Consider a population \( B \) that was isolated from a base population \( A \) some \( t_{AB} \) generations ago. Assume that \( B \) was split into two lineages, of which one underwent \( t_{BX} \) generations of drift leading to terminal taxon \( X \) and the other drifted for \( t_{BY} \) generations leading to terminal taxon \( Y \) (see fig. 1). We can observe the gene frequencies of \( X \) and \( Y \); can we infer the lengths of the three branches based on molecular data collected from \( X \) and \( Y \)? First, from the reduction in heterozygosity in \( X \) we can infer \( t_{AB} + t_{BX} \) based on \( H_X = H_A \gamma^{t_{BX}/t_{AB}} \), where \( H_A \) is the heterozygosity of population \( A \) and its remote ancestor \( A \), respectively (\( H_A \) is usually denoted by \( H_0 \)). Likewise, \( t_{AB} + t_{BY} \) can be inferred from \( H_Y = H_A \gamma^{t_{BY}/t_{AB}} \), where \( H_Y \) is the heterozygosity in \( Y \). Second, if we know the heterozygosity of the internal node, \( H_B \), we can infer \( t_{AB} \) using \( H_B = H_A \gamma^{t_{AB}/t_{AB}} \). If \( N_e \) and \( H_A \) are known, the above equations can be used to estimate the lengths of the three branches.

Heterozygosity and Genetic Distance

The estimated heterozygosity of terminal node \( X \) is given by

\[
D_{XX} = 1 - \sum_{i=1}^{n} x_i^2, \tag{2}
\]

where \( n \) is the number of allelic states of a particular locus, and \( x_i \) is the frequency of the \( i \)th allele for population \( X \). Under the assumption that \( X \) is in Hardy-Weinberg equilibrium, the expectation of \( D_{XX} \) equals \( H_X \). The heterozygosity of a terminal taxon can be calculated by sampling molecular data. However, heterozygosity of an internal node cannot be so calculated. We have previously shown that heterozygosity of the population at the internal node (e.g., \( B \)) can be estimated by (Xu, Atchley, and Fitch 1994)

\[
D_{XY} = 1 - \sum_{i=1}^{n} x_i y_i, \tag{3}
\]

where \( y_i \) is the frequency of the \( i \)th allele for population \( Y \), and \( \theta_{XY} \) approximately equals the average inbreeding coefficient of node \( B \), we have

\[
E(D_{XY}) = H_A(1 - \theta_{XY}), \tag{4}
\]

where \( F_B \) is the inbreeding coefficient of node \( B \). Although equation (3) is strictly an estimate of the heterozygosity of the population two generations after the split from node \( B \), we follow the convention adopted by Xu, Atchley, and Fitch (1994) and treat it as an estimate of node \( B \) itself. For population sizes normally observed in natural populations, the reduction of heterozygosity over two generations is negligible, and thus \( H_B \) can be approximated by \( D_{XY} \). With multiple loci, the overall estimated heterozygosity \( D_{XY} \) takes the average of all locus-specific genetic distances. Note that under the pure drift model, \( E(D_{XY}) \) is a function of \( t_{AB} \) but not a function of \( t_{BX} \) and \( t_{BY} \) (see fig. 1).

Reticulation Under Drift

Figure 2 shows that after lineage divergence, hybridization occurred at point \( C \) in the lineage leading to \( X \) and at point \( D \) in the lineage leading to \( Y \). The hybrid is denoted by node \( E \), and the terminal node of the hybrid lineage is denoted by \( Z \). Here, we assume that the hybrid contains equal contributions from the two progenitors. We also assume that the hybrid can reproduce true and show no change in fitness compared with its progenitors. The hybrid lineage evolves independent of

---

**Fig. 1.**—The simple rooted phylogeny used as an example in the text, where \( A \) is the base population, \( B \) is an internal node, \( X \) and \( Y \) are two terminal populations, and \( t_{ij} \) is the length of the branch between nodes \( i \) and \( j \).

**Fig. 2.**—The simple rooted reticulate phylogeny used as an example in the text, where a group of individuals from node \( C \) of lineage \( X \) hybridized with a group of individuals from node \( D \) of lineage \( Y \) to form a hybridized lineage \( Z \).
its parents (without backcrossing). Some of these assumptions may be relaxed (see Discussion).

It should be noted that hybridization between lineages \( X \) and \( Y \) in the past does not affect the genetic distance between \( X \) and \( Y \) (assuming no horizontal gene transfer). This allows the heterozygosity of node \( B \) to still be estimated by \( D_{XY} \). Figure 3 gives the reticulate phylogeny where we assume two hypothetical lineages diverged at the time when hybridization occurred. One lineage was isolated from node \( C \) and designated taxon \( X' \), while the other lineage was isolated from node \( D \) and designated taxon \( Y' \). Recall that the genetic distance between \( X \) and \( Z \) is defined as the probability that a random gene sampled from \( Z \) and a random gene from \( X \) have different allelic states. The random gene drawn from \( Z \) has a probability of \( 1/2 \) of coming from lineage \( Y \) and a probability of \( 1/2 \) of coming from lineage \( X \). If the sampled gene comes from lineage \( Y \), then \( D_{XZ} = D_{XY} \); otherwise, \( D_{XZ} = D_{XX} \), where \( D_{XX} \) is the genetic distance between \( X \) and \( Z \) and the hypothetical taxon \( X' \). Therefore, \( D_{XZ} = (D_{XX} + D_{XY})/2 \). Similarly, \( D_{YZ} = (D_{YY} + D_{XY})/2 \). From these two equations, we obtain

\[
D_{XX} = 2D_{XZ} - D_{XY}
\]

and

\[
D_{YY} = 2D_{YZ} - D_{XY}.
\]

These two genetic distances are important because \( D_{XX} \) and \( D_{YY} \) reflect the heterozygosities of internal nodes \( C \) and \( D \), respectively, which are required for estimating \( t_{BC} \) and \( t_{BD} \).

**Recovery of Heterozygosity in the Hybrid Lineage**

The hybrid population will regain some heterozygosity compared with its parents. Denote the first generation hybrid by \( F_1 \) and the second generation by \( F_2 \) (node \( E \) in fig. 3). The expected heterozygosity of \( F_1 \) is the probability that a random gene drawn from node \( C \) and a random gene from node \( D \) have different allelic states. As discussed earlier, this probability is the heterozygosity of the population two generations after node \( B \). Therefore, the heterozygosity of the \( F_1 \) hybrids is immediately recovered to the level at the time that the split just occurred. However, much of the regained heterozygosity will be lost in the next generation (\( F_2 \)). The amount of loss depends on \( H_C \) and \( H_D \). Xu and Atchley (1995) showed that the expected heterozygosity in \( F_2 \) (node \( E \)) is

\[
H_E = H_B/2 + (H_C + H_D)/4,
\]

provided that the effective population size is not too small, e.g., \( N_e > 50 \). If the hybridization occurs long after divergence between \( C \) and \( D \), then \( H_C \) and \( H_D \) will be small, leading to \( H_E = H_B/2 \); i.e., half of the heterozygosity will be lost. On the other hand, if hybridization occurs shortly after divergence between \( C \) and \( D \), then \( H_C \) and \( H_D \) will be close to \( H_B \), which leads to \( H_E = H_B \); i.e., there is no loss in heterozygosity.

As shown earlier, \( H_B, H_C \), and \( H_D \) can be estimated by \( D_{XY}, D_{XX} \), and \( D_{YY} \), respectively. The heterozygosity of node \( E \) can be estimated as

\[
H_E = D_{XY}/2 + (D_{XX} + D_{YY})/4.
\]

The heterozygosity in subsequent generations will decline at the usual rate, i.e., \( H_Z = H_E \gamma^{t_{EZ}} \).

**Expectation of Genetic Distance**

The genetic distances between nonhybrid species \( X \) and \( Y \) have the following expectations: \( E(D_{XX}) = H_C = H_Y = \gamma^{t_{AB} + t_{BC} + t_{CE}} \) and \( E(D_{YY}) = H_B = H_X = \gamma^{t_{AB} + t_{BD} + t_{DE}} \). The genetic distance of the hybrid lineage with itself has an expectation of:

\[
E(D_{ZZ}) = H_Z = H_E \gamma^{t_{EZ}} = (H_B/2 + (H_C + H_D)/4) \gamma^{t_{EZ}}
\]

\[
= H_A \gamma^{t_{EZ}} \left( \frac{1}{2} \gamma^{t_{AB}} + \frac{1}{4} \gamma^{t_{AB} + t_{BC}} + \frac{1}{4} \gamma^{t_{AB} + t_{BD}} \right)
\]

Note that by employing the approximation

\[
(1 - d)^{y} \approx 1 - dy
\]

for small \( d \), the terms in the parentheses of the above equation can be approximated by \( \gamma^{t_{AB} + t_{BC} + t_{BD}}/4 \) for large \( N_e \). Therefore,

\[
E(D_{ZZ}) = H_z = H_A \gamma^{t_{AB} + t_{BC} + t_{BD} + t_{DE}}
\]

The approximation is quite good even if \( N_e \) is small, as shown in figure 4.

Expectations of the genetic distances of the hybrid lineage from its parental species are

\[
E(D_{XZ}) = (H_B + H_C)/2 = H_A \left( \frac{1}{2} \gamma^{t_{AB}} + \frac{1}{4} \gamma^{t_{AB} + t_{BC}} \right)
\]

\[
\approx H_A \gamma^{t_{AB} + t_{BC} + t_{DE}}/2
\]

and

\[
E(D_{YZ}) = (H_B + H_D)/2 = H_A \left( \frac{1}{2} \gamma^{t_{AB}} + \frac{1}{4} \gamma^{t_{AB} + t_{BD}} \right)
\]

\[
\approx H_A \gamma^{t_{AB} + t_{BD} + t_{DE}}/2.
\]
Inferring Reticulate Phylogenies

The method is demonstrated using a hypothetical reticulate phylogeny with four taxa (see fig. 5). This phylogeny has a total of 8 branches (including the root) and 10 data points, allowing a least-squares method to be used to estimate the lengths of the branches.

First, define

\[ y_{ij} = [\ln(D_{ij}) - \ln(H_A)]/\ln \gamma, \]  

(14)

where \( i \) and \( j \) represent the terminal taxa. When \( N_e \) is large, \( \ln \gamma \approx -1/(2N_e) \), leading to

\[ y_{ij} = -2N_e \ln(D_{ij}) + 2N_e \ln(H_A). \]  

(15)

Note that \( H_A \) can be set at an arbitrary value without affecting the lengths of the tree branches except the length of the root \( (t_{AB}) \). The value of \( N_e \) is normally unknown, but the choice of \( N_e \) does not affect the relative relationships of the estimated branch lengths. Because \( H_A \) and \( N_e \) are common to all \( y_{ij} \), they can be ignored in the data analysis without affecting the estimation of relative branch lengths. Therefore, we take \( y_{ij} = -\ln(D_{ij}) \), which can be interpreted as a measure of genetic similarity between taxa \( i \) and \( j \). This similarity reflects the time during which taxa \( i \) and \( j \) share the same evolutionary pathways.

Define \( \mathbf{y} = \{y_{XX}, y_{YW}, y_{YZ}, y_{YY}, y_{XZ}, y_{XY}, y_{WZ}, y_{WX}, y_{YZ}\}^T \) as a vector of data, \( \mathbf{t} = \{t_{AB}, t_{BC}, t_{CF}, t_{FX}, t_{FW}, t_{DY}, t_{Z}\}^T \) as the lengths of branches of the phylogeny, and \( \mathbf{e} = \{e_{XX}, e_{YW}, e_{ZZ}, e_{YY}, e_{XZ}, e_{XY}, e_{WZ}, e_{WX}, e_{YZ}\}^T \) as a vector of residual errors. We have the following linear model:

\[ \mathbf{y} = \mathbf{Xt} + \mathbf{e}, \]

(16)

where \( \mathbf{X} \) is the design matrix determined by the tree topology

\[
\begin{bmatrix}
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

(17)

The least-squares estimate of \( \mathbf{t} \) is

\[ \hat{\mathbf{t}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \]

(18)

with a mean squared error of

\[ \text{MSE} = \frac{1}{df}(\mathbf{y} - \hat{\mathbf{X}}^T\hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{X}}^T\hat{\mathbf{y}}), \]

(19)

where degrees of freedom \( (df) = 10 - 8 = 2 \). Note that one degree of freedom has been lost compared with a regular bifurcating tree.

The total number of possible reticulate phylogenies may be huge for a large number of taxa. If it is known a priori that some taxa are hybrids and the hybridization has occurred at most once in any given lineage, the total number of reticulate phylogenies may be significantly reduced such that it is possible to search for the best phylogeny (with the least MSE). Consider, for example, the four-taxon reticulate evolution model. If we know that taxon \( Z \) has a hybrid origin but we do not know where the hybridization occurred, then the total number of reticulate phylogenies is 12 (see fig. 6). If we know that one of the four taxa is a hybrid but we do not know which one, then the total number of possible phylogenies will increase to \( 4 \times 12 = 48 \). Theoretically, all possible phylogenies must be evaluated, and the inferred phylogeny is the one that has the minimum MSE.

The Mutation Model

The mutation model was developed according to the theory formulated by Cockerham (1984), whereby
Performing log transformation, we have

\[\log(q^* - q_{XY}) = \log(q^* - Q^*\alpha^{t_{XY}+t_D})\]

**Equation (20),** we have \((q^* - q_{XY}) = (q^* - Q^*)\alpha^{t_{XY}+t_D}\), which leads to

\[y_{XY} = \frac{\ln(q^* - q_{XY}) - \ln(q^* - Q^*)}{\ln \alpha} = t_{BX} + t_{BY}\]  

**Equation (21)**

In contrast to the drift model, in which \(y_{XY}\) is proportional to the time before \(X\) and \(Y\) split, the \(y_{XY}\) under the mutation model is proportional to the time after \(X\) and \(Y\) diverged.

I now examine the gene alike indices in the reticulate phylogeny (fig. 2). The internal node of the hybrid can be decomposed into \(E = \delta C + (1 - \delta)D\), where \(\delta\) is an indicator variable defined as \(\delta = 1\) if an allele sampled from \(Z\) comes from \(C\) and \(\delta = 0\) if the allele comes from \(D\). With equal contributions from \(C\) and \(D\), we have \(\Pr(\delta = 1) = 1/2\). Therefore,

\[q^* - q_{XY} = (q^* - Q^*)\alpha^{t_{CX}+t_{EZ} + \frac{1}{2}t_{CX}+t_{BC}+t_{BD}+t_{EZ}}\]

This approximation is made under the assumption that \(u + v < 1\). Similarly,

\[q^* - q_{YZ} = (q^* - Q^*)\alpha^{t_{BY}+t_{BD} + \frac{1}{2}t_{BY}+t_{BC}+t_{BD}+t_{EZ}}\]  

**Equation (22)**

Performing log transformation, we have

\[\log(q^* - q_{XY}) = \log(q^* - Q^*)\alpha^{t_{XY}+t_D}\]

**Equation (20)**

the number of allelic states per locus \((n)\) is assumed to be finite. For simplicity, I also assumed an equal mutation rate \((v)\) of any allele to any other specific allele so that the overall mutation rate is \(u = (n - 1)v\). Cockerham’s (1984) gene alike index is used for the derivation. The gene alike index within population \(X\) is defined as the probability of a random pair of alleles being alike (identical by state) and is estimated by \(Q = 1 - D_{XX} = \sum_{i=1}^{n} x_i^2\). Similarly, the gene alike index between populations \(X\) and \(Y\) is defined as \(q_{XY} = 1 - D_{XY} = \sum_{i=1}^{n} x_i Y_i\). The phylogeny is assumed to have started with an equilibrium gene alike value in the ancestral population before the taxa diverged. This assumption implies that the gene alike index within each population and each internal node is a constant; i.e., \(Q = Q^*\) for all \(i\)’s, so that only \(q_{XY}\) is informative for phylogenetic inference. Cockerham (1984) provided the equilibrium gene alike value \(Q^* = (1 + 4n_v)/(1 + 4n_nv)\). The assumption of equilibrium \(Q^*_X\) does not imply constant allelic frequencies over time. In fact, the allelic frequencies must change from time to time so that the gene alike index between populations also varies overtime.

Consider the simple phylogeny given in figure 1. The index of gene alike between \(X\) and \(Y\) is

\[q_{XY} = q^* - (q^* - Q^*)\alpha^{t_{XY}+t_D}\]  

**Equation (20)**

where \(\alpha = 1 - u - v\), and \(q^* = 1/n\) is the equilibrium value of \(q_{XY}\) (Cockerham 1984). Rearranging equation

**FIG. 6.—**The 12 possible rooted reticulate phylogenies for four taxa with \(Z\) as the hybrid.
The same least-squares method is applied here to evaluate the tree and estimate the lengths of the branches.

**Numerical Studies**

The reticulate phylogeny given in figure 5 was simulated under the pure drift model. First, a base population (node A) was simulated, from which 50 males and 50 females were randomly sampled, forming parents of the next generation. The population then underwent 50 generations of random mating and random selection, creating node B, which was then split into two lineages. One lineage diverged to node C after 50 generations of drift, and the other lineage diverged to node D. Node C continued drifting for another 50 generations into node E, which was itself split into two populations, X and W; each underwent 50 generations of drift. Node D continued drifting for 100 generations, leading to Y. Beyond the normal bifurcating process, a hybrid population (node E) was formed via hybridization between 50 males from node C and 50 females from node D. The hybrids then underwent random drift for 100 generations, leading to the terminal taxon Z.

Let $D$ be the distance matrix with the taxa arranged in the order of \{X W Z Y\}, i.e.,

$$D = \begin{bmatrix}
D_{XX} & D_{XY} & D_{YZ} \\
D_{XY} & D_{YY} & D_{ZZ} \\
D_{XY} & D_{YY} & D_{ZZ}
\end{bmatrix}.$$ 

Under this setting, the effective population size is $N_e = 100$. Let $H_A = 0.50$; then, the expectation of $D$ can be obtained using equations (11)–(13):

$$E(D) = \begin{bmatrix}
0.1835 & 0.2357 & 0.3460 & 0.3891 \\
0.2357 & 0.1835 & 0.3460 & 0.3891 \\
0.3460 & 0.3460 & 0.2096 & 0.3460 \\
0.3891 & 0.3891 & 0.3460 & 0.1835
\end{bmatrix}.$$ 

$$= \begin{bmatrix}
0.1835 & 0.2357 & 0.3433 & 0.3891 \\
0.2357 & 0.1835 & 0.3433 & 0.3891 \\
0.3433 & 0.3433 & 0.2079 & 0.3433 \\
0.3891 & 0.3891 & 0.3433 & 0.1835
\end{bmatrix}.$$ 

Note that the differences between the exact and the approximated $D$ are negligible.

The reticulate phylogeny (fig. 5) was then simulated under 100 independent biallelic, equally frequent loci. The genetic distances calculated from a single run of the simulation were

$$D = \begin{bmatrix}
0.1843 & 0.2315 & 0.3188 & 0.3580 \\
0.2315 & 0.1684 & 0.3088 & 0.3615 \\
0.3188 & 0.3088 & 0.1839 & 0.3157 \\
0.3580 & 0.3615 & 0.3157 & 0.1787
\end{bmatrix}.$$ 

which are reasonably close to $E(D)$.

The genetic distances are converted into $y$ using equation (14) under $H_A = 0.50$ and $N_e = 100$. The data are then fitted to the linear model with $X$ chosen under the true phylogeny, given in equation (17) (see fig. 5). The lengths of branches are estimated using equation (18), with results also shown in figure 5 (values with floating points). In general, the estimated lengths of branches are reasonably close to the true values, although some sampling errors have been observed. The MSE value of this phylogeny is 11.05 generation$^2$.

The data ($y$) are then fitted to each of the remaining 11 reticulate phylogenies (fig. 6) and the 15 bifurcating trees (fig. 7). The MSEs of these trees are given in table 1, showing that the true reticulate phylogeny (tree 2) does have the least MSE.

Repeated simulations were also conducted under the same setting. The number of loci ($L$), however, was examined at the following levels: $L = 10, 20, 30, 50, 75, 100, 150, 200$. The simulated molecular data were analyzed under each of the 12 possible reticulate phy-
Table 1
Mean Squared Errors (MSEs) of the 12 Possible Reticulate Phylogenies with Taxon Z Being the Hybrid (see fig. 6) and the 15 Possible Bifurcating Trees (see fig. 7) Calculated from a Single Simulation with 100 Independent Loci

<table>
<thead>
<tr>
<th>Phylogeny ID</th>
<th>MSE</th>
<th>Phylogeny ID</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,054.26</td>
<td>1 or 2 or 3</td>
<td>2,774.22</td>
</tr>
<tr>
<td>2</td>
<td>11.05</td>
<td>4 or 7 or 11</td>
<td>3,331.41</td>
</tr>
<tr>
<td>3</td>
<td>316.22</td>
<td>2 or 11</td>
<td>5,574.89</td>
</tr>
<tr>
<td>4</td>
<td>754.61</td>
<td>3 or 11</td>
<td>5,685.21</td>
</tr>
<tr>
<td>5</td>
<td>6,404.97</td>
<td>6 or 11</td>
<td>6,404.97</td>
</tr>
<tr>
<td>6</td>
<td>5,574.89</td>
<td>5 or 11</td>
<td>6,404.97</td>
</tr>
<tr>
<td>7</td>
<td>5,685.21</td>
<td>7 or 11</td>
<td>2,662.68</td>
</tr>
<tr>
<td>8</td>
<td>2,662.68</td>
<td>8 or 11</td>
<td>1,933.95</td>
</tr>
<tr>
<td>9</td>
<td>6,509.13</td>
<td>9 or 11</td>
<td>2,937.93</td>
</tr>
<tr>
<td>10</td>
<td>6,509.13</td>
<td>10 or 11</td>
<td>1,452.29</td>
</tr>
<tr>
<td>11</td>
<td>6,725.81</td>
<td>11 or 11</td>
<td>3,011.08</td>
</tr>
<tr>
<td>12</td>
<td>6,725.81</td>
<td>12 or 11</td>
<td>1,484.01</td>
</tr>
<tr>
<td>13</td>
<td>5,748.97</td>
<td>13 or 11</td>
<td>2,894.95</td>
</tr>
<tr>
<td>14</td>
<td>5,748.97</td>
<td>14 or 11</td>
<td>2,860.39</td>
</tr>
<tr>
<td>15</td>
<td>5,748.97</td>
<td>15 or 11</td>
<td>2,255.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 7.—The 15 possible rooted bifurcating trees for four taxa.

NOTE.—Reticulate phylogeny 2 is the true one.

logenies (fig. 6) and the 15 possible bifurcating trees (fig. 7). The inferred phylogeny is chosen as the one that has the minimum MSE among the 27 phylogenies considered. The simulation was then repeated 100 times for each level of \( L \).

The three topologies simulated represent three different cases: hybridization between closely related taxa, ancient hybridization, and new hybridization between distantly related taxa. The first simulated topology is reticulate phylogeny 1 (see fig. 6), where Z is a hybrid between X and W, which are sister taxa (recently diverged). The frequency of being chosen as the inferred phylogeny is given in table 2 for each of the 12 + 15 = 27 phylogenies considered. When the number of loci is small, the true phylogeny has a low frequency of being inferred compared with bifurcating trees 6 and 10. Note that the relationship between the hybrid and its progenitors is \((X, Z), W\) in tree 6, while the relationship is \((X, Z, W)\) in tree 10. As the number of loci increases, the frequency of the true phylogeny begins to dominate.

The second topology simulated is reticulate phylogeny 2 (see fig. 6), where Z is a hybrid between clade \((X, W)\) and \(Y\), which are distantly related. The hybrid-
ization occurred shortly after the two lineages diverged (old hybridization). The frequency of being chosen as the inferred phylogeny is given in Table 3 for each possible phylogeny. There are 12 possible reticulate phylogenies and 15 possible bifurcating trees. Reticulate phylogeny number 1 is the true phylogeny.

Discussion

I have developed a method to analyze reticulate phylogenies under the pure drift model and subsequently extended the method to fit the mutation model. The drift model is best used in short-term evolution, in which mutation has not played an important role. The mutation model, however, is applicable for long-term evolution, in which the within-population heterozygosity of any internal node has reached an equilibrium. Both models use genetic distances (or similarities) between taxa in gene frequency. However, the genetic distance between populations under the drift model is a function of the time (t) during which the two populations share the same evolutionary pathways, whereas the genetic distance under the mutation model is a function of the time (t) since the two populations have diverged. While the latter is easily understood, the former appears to be counterintuitive. Under the pure drift model, the genetic distance between two populations reflects the heterozygosity of the immediate common ancestor and contains no information about the heterozygosity of either population. The heterozygosity of the immediate common

Table 2
Simulation Results from a Reticulate Phylogeny in Which the Hybrid Species is Derived from Recently Diverged Sister Taxa

<table>
<thead>
<tr>
<th>PHYLOGENY ID</th>
<th>NO. OF LOCI</th>
<th>RETICULATION</th>
<th>BIFURCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>150</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note.—The table entries are the frequencies (out of 100 replicates) of being chosen as the inferred phylogeny for each possible phylogeny. There are 12 possible reticulate phylogenies and 15 possible bifurcating trees. Reticulate phylogeny number 1 is the true phylogeny.

Table 3
Simulation Results from a Reticulate Phylogeny in Which the Hybrid Species is Derived from Distantly Related Taxa but Arose Only Shortly After These Lineages Diverged

<table>
<thead>
<tr>
<th>PHYLOGENY ID</th>
<th>NO. OF LOCI</th>
<th>RETICULATION</th>
<th>BIFURCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>150</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note.—The table entries are the frequencies (out of 100 replicates) of being chosen as the inferred phylogeny for each possible phylogeny. There are 12 possible reticulate phylogenies and 15 possible bifurcating trees. Reticulate phylogeny number 2 is the true phylogeny.
Bifurcation

logeny given in mutation are important. Unfortunately, derivation of the
be combined to infer phylogenies when both drift and
of the two populations). In theory, the two models can
will work, because the genetic distance measured this
by the joint force of drift and mutation, neither model
sive. In situations in which the phylogeny was shaped
ancestors after the branch lengths
common ancestor was isolated from the base pop-
ancestor, however, is only a function of the time (t) since
common ancestor was isolated from the base pop-
ulation (the root of the phylogeny). The segments
between the base population and the immediate common
ancestors are the common pathways shared by the two
populations before their divergence.

The drift model and the mutation model are exclu-
sive. In situations in which the phylogeny was shaped
by the joint force of drift and mutation, neither model
will work, because the genetic distance measured this
way is a function of both r’s (before and after the split
of the two populations). In theory, the two models can
be combined to infer phylogenies when both drift and
mutation are important. Unfortunately, derivation of the
drift-mutation model is complicated. Consider the
phylogeny given in figure 1. The within-population gene
alike indices are (Cockerham 1984)

\[ Q^* - Q_X = (Q^* - Q_B)\gamma^{fx} = (Q^* - Q_A)\gamma^{AB+fx}, \]

\[ Q^* - Q_Y = (Q^* - Q_B)\gamma^{fy} = (Q^* - Q_A)\gamma^{AB+fy}, \]

and the between-population index is

\[ q^* - q_{XY} = (q^* - Q_B)\alpha^{fx+fy}, \]

which can be re-expressed as

\[ (Q^* - q_{XY}) = [(q^* - Q^*) + (Q^* - Q_B)]\alpha^{fx+fy} \]

\[ - (q^* - Q^*) \]

\[ = (Q^* - Q_A)\gamma^{AB}\alpha^{fx+fy} \]

\[ + (q^* - Q^*)[\alpha^{fx+fy} - 1]. \]

Note that \( Q^* - q_{XY} \) is a function of both \( t_{AB} \) (the
time before the populations split) and \( t_{fx} + t_{fy} \) (the
time after the populations split). The complexity comes
from the fact that the log of \( Q^* - q_{XY} \) is not a linear function
of the \( t \)’s unless

\[ (q^* - Q^*)[\alpha^{fx+fy} - 1] = 0. \]

Unfortunately, equation (30) does not hold in general.
Even if it holds, we can see that

\[ y_{XY} = \frac{\ln(Q^* - q_{XY}) - \ln(Q^* - Q_A)}{\ln \gamma} \]

\[ = \frac{t_{AB} + (t_{fx} + t_{fy})\ln \alpha}{\ln \gamma}, \]

which is still complicated because the branch lengths
before and after the populations split are expressed in
different scales. Alternative methods, such as parsimony
and maximum likelihood, may be preferable for com-
bining the two models; this possibility deserves further
investigation.

Molecular data at the sequence level have been
used to detect horizontal gene transfer, e.g., recombi-
nation within a sequence (Hein 1990, 1993; Hudson
1990; Bolyky et al. 1996; Grassy and Holmes 1997).
It is not clear how useful it is to infer species hybrid-
ization using sequence data. Under certain circumstanc-
es, nuclear DNA polymorphism in restriction endonu-
cleases may be used to infer reticulate phylogenies. One
of the basic assumptions when using restriction data is
that the sites must be independent. This assumption
may hold when the restriction sites are located far apart
on the genome such that the sites freely recombine in
the hybrid lineage. Alternatively, if the hybrid lineage is
formed by the hybridization of a large number of indi-
viduals from each parental lineage, given a sufficient
number of generations in random mating within the hy-
drid lineage, the sites may behave as independent, even
if they are located close together. Nei and Li (1979)
developed the mathematical model for studying popu-
lation divergence in terms of restriction endonucleases.
The proportion of sites shared by lineages \( X \) and \( Y \), de-
noted by \( S_{XY} \), is expected to decline as \( X \) and \( Y \) further
diverge. Nei and Li (1979) showed that

\[ S_{XY} = e^{-\lambda(t_{fx}+t_{fy})}, \]

where \( r \) is the number of nucleotides in the restriction
site (e.g., \( r = 6 \) for EcoRI-GAATTC), \( \lambda \) is the rate of
nucleotide substitution per unit time (year or genera-
tion), and \( t_{fx} + t_{fy} \) is the total amount of time since \( X \)
and \( Y \) diverged (see fig. 1). When multiple restriction
enzymes are used, the sites from the same \( r \)-valued re-
striction enzymes should be combined. Data from different \( r \)-valued restriction enzymes should be properly weighted before being used for analysis. Define \( S_{XY} \), as the proportion of shared sites between \( X \) and \( Y \) for an enzyme with \( r_j \) nucleotides in the restriction site. The following combination of data is suggested:

\[
y_{XY} = -\sum_{i=1}^{m} r_i^{-1} \ln S_{XY_i} = \lambda (t_{BX} + t_{BY}),
\]

where \( m \) is the number of different restriction enzymes used. The distance between \( X \) and \( Y \) is finally expressed as a linear function of the times after they diverged. The distance involving a hybrid lineage can be similarly expressed, e.g.,

\[
y_{XZ} = \lambda \left( t_{CX} + \frac{1}{2} t_{BE} + \frac{1}{2} t_{BD} + t_{EZ} \right)
\]

The same least-squares method can be used to evaluate a reticulate phylogeny (see the mutation model in gene frequency).

Recombination, a form of reticulation at the gene level, generates the same problems as hybridization. Methods exist which try to diagnose recombination by looking at the compatibility of the "phylogenetic partition" supported by the polymorphic sites along the sequence (Drouin and Dover 1990), by looking at changes in the most parsimonious topology along sequences (Hein 1990, 1993), by using a maximum chi-square test (Maynard Smith 1992), or by using the maximum-likelihood approach to detect the specific region showing "anomalous" evolutionary patterns (Grassly and Holmes 1997). However, no general methods exist which allow the placement of a putative hybrid in the correct position relative to the two progenitors if the progenitors are unknown, it does not allow the placement of the hybrid in the correct position relative to the two progenitors if the progenitors are known. The above theoretical works have enhanced our understanding of reticulate evolution, but they may only represent a small proportion of the work required to complete a more general approach.

To obtain the number of loci required for this method to work accurately, dominant markers such as AFLPs may have to be employed. The method proposed can handle dominant markers provided that the gene frequency within a population can be estimated using the Hardy-Weinberg law; i.e., the frequency of the recessive allele is estimated by the square root of the frequency of the recessive homozygotes. The number of allelic states per dominant locus is considered to be two. The efficiency using dominant markers would be slightly less than that observed in the biallelic codominant system (see the simulation studies section) because the gene frequency is not given, but estimated from the genotypic frequencies.

The pure drift model may be of interest in its own right. Inbred strains of laboratory animals are valuable model organisms for studies in evolutionary biology, particularly at the molecular level (see Atchley and Fitch 1991, 1993; Fitch and Atchley 1985, 1989). The phylogeny of inbred strains is most likely driven by genetic drift, not by mutation. First, most of the inbred strains of mice could have arisen from just a few mice (Atchley and Fitch 1993). Second, most of the inbred strains are derived by systematic brother × sister mating, which represents the maximum effect of genetic drift in laboratory animals. Third, the evolutionary history of these organisms is too short for a significant mutational input (most laboratory strains of rats and mice have been inbred for less than 200 generations). However, many inbred mouse and rat strains were originally produced from hybridization between genetically divergent strains (Atchley and Fitch 1993). For instance, the SEC strain of mice was derived from hybrids between NB and BALB/c (Festing 1989), and the BS strain of rats was derived from hybrids between NZ and a wild rat (Hedrich 1990). With the theory presented in this paper and the data from inbred strains of animals with known hybrid origins, the genetic aspects of reticulation and its impact on phylogenetic inference could be studied in detail. Furthermore, the model may be readily applied to evolutionary studies of domesticated animals and agricultural cultivars.

For generality, suppose that backcrosses occurred a few times immediately after the initial hybridization event, such that the hybrid taxon ultimately inherits a proportion \( p \) of genes from parental taxon \( X \) and a proportion \( 1 - p \) of genes from parental taxon \( Y \) (see fig. 2). In this case, the expected genetic distances become

\[
E(D_{XY}) = pE(D_{XY}) + (1 - p)E(D_{XZ})
\]

\[
= pH_c + (1 - p)H_B \quad \text{and} \quad E(D_{YX}) = (1 - p)E(D_{YY}) + pE(D_{XY})
\]

\[
= (1 - p)H_D + pH_B.
\]

Estimation of branch lengths under this unidirectional gene flow scenario is still possible if \( p \) is known. Otherwise, data from more taxa are required to estimate branch lengths and \( p \) simultaneously.

A final caveat concerns the assumption of constant effective population size along all segments of the phylogeny. This assumption is not realistic, especially for the hybridized lineage. In the early stage of hybridization, the hybrid population must have experienced a sort of bottleneck and selection. There may be much reorganization of the genome, linkage disequilibrium of genes or chromosomal blocks, as nicely demonstrated in an empirical case (Rieseberg, Vanfossen, and Desrochers 1995). The robustness of the model to these effects needs to be further studied. Nonetheless, we can slightly relax the assumption of constant \( N_e \) by assuming that \( N_e \) is constant within a segment, but it can vary across different segments. In this case, the estimated branch length for each segment is the number of generations divided by twice the effective population size corresponding to that period of time. A similar argument
also holds for the assumption of constant mutation rate $v$ across loci and across alleles within loci.

Acknowledgments

I thank Damian D. Gessler for justifying the approximations in equations (11)–(13) and for his help in smoothing the presentation. I am grateful to Norman Ellstrand, Richard Whitkus, and Claus Vogl for comments on an earlier version of the manuscript. This work was partly supported by National Institutes of Health grant GM55321.

LITERATURE CITED


YUN-XIN FU, reviewing editor

Accepted February 10, 2000